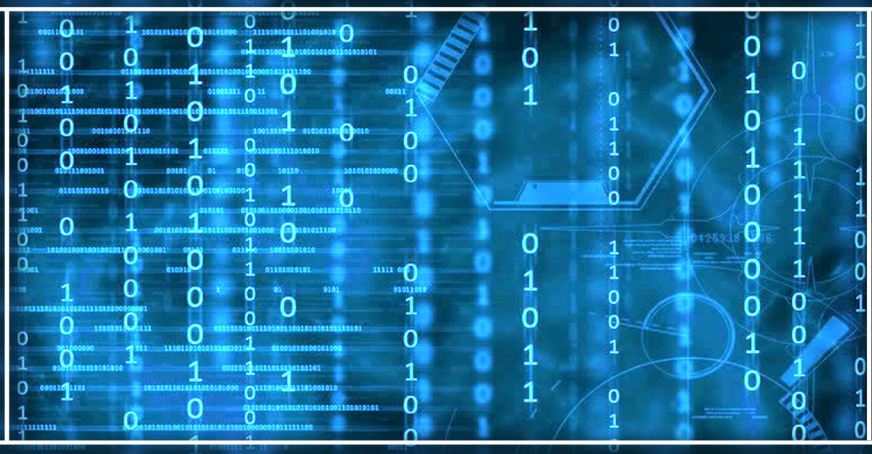
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 16 Issue 10
October 2025



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

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

Thank you for Sharing Wisdom!

Kohei Arai
Editor-in-Chief
IJACSA
Volume 16 Issue 10 October 2025
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Editor-in-Chief

Dr. Kohei Arai - Saga University
Domains of Research: Artificial Intelligence, Big Data Analytics, Computer Vision, Decision Making, Information Retrieval, Satellite Remote Sensing

---

Associate Editors

Alaa Sheta
Southern Connecticut State University
Domain of Research: Artificial Neural Networks, Computer Vision, Image Processing, Neural Networks, Neuro-Fuzzy Systems

Arun Kulkarni
University of Texas at Tyler
Domain of Research: Machine Vision, Artificial Intelligence, Computer Vision, Data Mining, Image Processing, Machine Learning, Machine Vision, Neural Networks, Neuro-Fuzzy Systems

Dr Ronak Al-Haddad
Anglia Ruskin University / Cambridge
Domain of Research: Technology Trends, Communication, Security, Software Engineering and Quality, Computer Networks, Cyber Security, Green Computing, Multimedia Communication, Network Security, Quality of Service

Elena Scutelnicu
"Dunarea de Jos" University of Galati
Domain of Research: e-Learning, e-Learning Tools, Simulation

In Soo Lee
Kyungpook National University
Domain of Research: Intelligent Systems, Artificial Neural Networks, Computational Intelligence, Neural Networks, Perception and Learning

Larbi Boubchir
University of Paris 8
Domain of Research: Artificial Intelligence, Biometrics

Renato De Leone
University of Camerino
Domain of Research: Mathematical Programming, Large-Scale Parallel Optimization, Transportation problems, Classification problems, Linear and Integer Programming

Xiao-Zhi Gao
University of Eastern Finland
Domain of Research: Artificial Intelligence, Genetic Algorithms

CONTENTS

Paper 1: AI-Assisted Workflow Optimization and Automation in the Compliance Technology Field
Authors: Zhen Zhong
PAGE 1 – 5

Paper 2: From Legacy to Cloud: Migration Strategies for Traditional Financial Institutions Using AWS
Authors: Uday Kiran Chilakalapalli, Brij Mohan, Vinodkumar Reddy Surasani
PAGE 6 – 15

Paper 3: From Logs to Knowledge: LLM-Powered Dynamic Knowledge Graphs for Real-Time Cloud Observability
Authors: Nurmyrat Amanmadov, Tarlan Abdullayev
PAGE 16 – 25

Paper 4: IoT-Enabled Data-Driven Optimization of Dynamic Thermal Loads for Low-Energy Buildings
Authors: Zhaojiang Lyu
PAGE 26 – 37

Paper 5: Quantifying Career Preferences and Perceptions of Software Testing Among Filipino IT Students: A Mixed-Method Analysis
Authors: Chrisza Joy M. Carrido, Abeer Alsadoon, Thair Al-Dala'in, Ahmed Hamza Osman, Abubakar Elsafi, Azhari Qismallah, Albaraa Abuobieda
PAGE 38 – 50

Paper 6: Pedestrian Navigation System with 3D Map and Charging Server Based on Steganography
Authors: Kohei Arai
PAGE 51 – 59

Paper 7: Comparative Evaluation of CNN Architectures for Skin Cancer Classification
Authors: Taopik Hidayat, Nurul Khasanah, Elly Firasari, Laela Kurniawati, Eni Heni Hermaliani
PAGE 60 – 68

Paper 8: TabNet–XGBoost Hybrid Model for Student Performance Prediction and Customized Feedback
Authors: Anupama Prasanth
PAGE 69 – 79

Paper 9: Enhanced Fault Detection in Software Using an Adaptive Neural Algorithm
Authors: Jasem Alostad
PAGE 80 – 89

Paper 10: Construction and Characteristics of an Engineering Economic Risk Management Platform Based on the BO-GBM Model
Authors: Chaojian Wang, Die Liu
PAGE 90 – 103

Paper 11: Adaptive Virtual Machine Consolidation Based on Autoformer and Enhanced Double Q-Network for Energy-Efficient Cloud Data Center
Authors: Kaiqi Zhang, Youbo Lyu, Dequan Zheng, Yanping Chen, Jianshan Xu
PAGE 104 – 119

Paper 12: Handwriting Detectives Using Wavelet Siamese Technology to Verify Signature Fraud
Authors: Mohamed Nazir, Ali Maher, Mostafa Eltokhy, Ali M. El-Rifaie, Tarek Hosny, Hani M. K. Mahdi
PAGE 120 – 130

Paper 13: A New Hybrid Algorithm for Vision-Based Sleep Posture Analysis Integrating CNN, LSTM and MediaPipe
Authors: Apichaya Nimkoompai, Puwadol Sirikongtham
PAGE 131 – 139

Paper 14: NetDAIL: An Optimized Deep Learning-Based Hybrid Model for Anomaly Detection in Network Traffic
Authors: Saad Khalifa, Mohamed Marie, Wael Mohamed
PAGE 140 – 150

Paper 15: An Integrated CNN, YOLOv5 and Faster R-CNN Framework for Real-Time Water Pipe Defect Detection
Authors: Chu Fu, Mideth Abisado
PAGE 151 – 157

Paper 16: SD-CNN: A Novel Lightweight Convolutional Neural Network Model for Fall Detection
Authors: Han-lin Shen, Tian-hu Wang, Hong Mu
PAGE 158 – 164

Paper 17: Efficient Lightweight Detection and Classification Method for Field-Grown Horticultural Crops
Authors: Yaru Huang, Hua Zhou, Zhongyi Shu
PAGE 165 – 176

Paper 18: Using Combined Weighting and BP Neural Networks for Relative Poverty Measurement and its Evaluation
Authors: Xiaohua Cai, Ya Zhao, Lijia Chen, Juan Huang, Yang Xu
PAGE 177 – 184

Paper 19: Evaluating Transformer-Based Pretrained Models for Classical Arabic Named Entity Recognition
Authors: Mariam Muhammed, Shahira Azab
PAGE 185 – 193

Paper 20: A Hippopotamus Optimization Algorithm-Based Convolutional Neural Network Model for Mental Health Assessment Among College Students
Authors: Gai Hang, Lin Yang
PAGE 194 – 202

Paper 21: A Hybrid Deep Learning and IoT Framework for Predictive Maintenance of Wind Turbines: Enhancing Reliability and Reducing Downtime
Authors: Amina Eljyidi, Hakim Jebari, Siham Rekiek, Kamal Reklaoui
PAGE 203 – 211

Paper 22: Federated Machine Learning for Monitoring Student Mental Health in Kazakhstan
Authors: Bakirova Gulnaz, Bektemyssova Gulnara, Nor'ashikin Binti Ali
PAGE 212 – 220

Paper 23: Exploring Hallucination in Large Language Models
Authors: Nesreen M. Alharbi, Thoria Alghamdi, Raghda M. Alqurashi, Reem Alwashmi, Amal Babour, Entisar Alkayal
PAGE 221 – 228

Paper 24: Game-Theoretic Approaches for Robust Stability of DC Motor Systems
Authors: Mohamed Ayari, Atef Gharbi, Yamen El Touati, Zeineb Klai, Mahmoud Salaheldin Elsayed, Elsaid Md. Abdelrahim
PAGE 229 – 236

Paper 25: Comparative Performance Analysis of Original AuRa and Improved AuRa Consensus Algorithms in Chain Hammer Digital Certificate Simulation
Authors: Robiah Arifin, Wan Aezwani Wan Abu Bakar, Mustafa Man, Evizal Abdul Kadir
PAGE 237 – 244

Paper 26: An Improved Marine Predators Algorithm-Based UAV Path Planning for 10-kV Distribution Networks Inspection in Live Working Scenarios
Authors: Dapeng Ma, Hongtao Jiang, Lichao Jiang, Chi Zhang, Changwu Li, Xin Zheng, Mingxian Liu, Kai Li
PAGE 245 – 254

Paper 27: Enhancing Predictive Maintenance Method Using Machine Learning to Improve IoT-Embedded Machinery Efficiency and Performance
Authors: Abiinesh Nadarajan, Iskandar Ishak, Noridayu Manshor, Raihani Mohamed, Mohamad Yusnisyahmi Yusof
PAGE 255 – 264

Paper 28: A Hybrid Deep Learning and Forensic Approach for Robust Deepfake Detection
Authors: Sales Aribe Jr
PAGE 265 – 275

Paper 29: Facial Expression Recognition Under Partial Occlusion Using Part-Based Ensemble Learning
Authors: Evangelions Felix Yehdeya, Wahyono
PAGE 276 – 284

Paper 30: Ambulance Detection and Priority Passage at Urban Intersections Using Transfer Learning and Explainable AI
Authors: Murtaza Hanif, Taj Muhammad, Atif Ikram, Shahid Yousaf, Marwan Abu-Zanona, Asef Mohammad Ali Al Khateeb, Bassam Elzaghmouri, Saad Mamoun Abdel Rahman Ahmed, Lamia Hassan Rahamatalla
PAGE 285 – 292

Paper 31: Digital Trust and Legacy: Mapping the Intersection of Inheritance Systems and Emerging Technologies (2010–2025)
Authors: Nor Aimuni Md Rashid, Faiqah Hafidzah Halim, Hazrati Zaini, Norshahidatul Hasana Ishak, Nur Farahin Mohd Johari, Alya Geogiana Buja
PAGE 293 – 300

Paper 32: A Quality Assessment Study of Deep Learning Techniques for Medical Image Diagnosis and Their Applications: A Systematic Literature Review
Authors: Amine Berquedich, Ahmed Zellou
PAGE 301 – 311

Paper 33: Modeling and Analyzing Malware Behavior in Virtual Networks Using EVE-NG
Authors: Maria-Mădălina Andronache, Alexandru Vulpe, Corneliu Burileanu
PAGE 312 – 327

Paper 34: Predicting Stock Market Performance Based on Sentiment Analysis of Online Comments
Authors: Wenhao Suo, Tongjai Yampaka
PAGE 328 – 339

Paper 35: User Satisfaction in AI-Driven Islamic Fintech: An Extended Technology Acceptance Model with Task–Technology Fit and Sharia Compliance

Authors: Mardiana Andarwati, Sari Yuniarti, Andriyan Rizki Jatmiko, Firnanda Al-Islama Achyunda Putra, Galandaru Swalaganata, Ahmad Taufiq Andriono

PAGE 340 – 349

Paper 36: Multimodal Deep Learning for Tuberculosis Detection Using Cough Audio and Clinical Data with Health Acoustic Representations (HeAR)

Authors: Rinaldi Anwar Buyung, Widi Nugroho

PAGE 350 – 357

Paper 37: EYE-GDM: Clinically Validated, Explainable Ensemble Learning for Gestational Diabetes

Authors: Shatha Alghamdi, Rashid Mehmood, Fahad Alqurashi, Turki Alghamdi, Sarah Ghazali, Asmaa AlAhmadi

PAGE 358 – 370

Paper 38: An Intelligent Platform for Behavior Modification and Office Syndrome Risk Reduction Using MediaPipe and Computer Vision

Authors: Sumran Chaikhamwang, Wijitra Montri, Chalida Janthajirakowit, Srinuan fongmanee

PAGE 371 – 379

Paper 39: A Review of Ransomware Detection Models for Cybersecurity Driven IIoT in Cloud Environments

Authors: Abrar Ali, Norah Hamed, Monir Abdullah

PAGE 380 – 389

Paper 40: A Comprehensive Survey of Visual SLAM Technology: Methods, Challenges, and Perspectives

Authors: Aidos Ibrayev, Amanzhol Bektemessov

PAGE 390 – 400

Paper 41: Uneven But Accelerating: AI Adoption in Higher Education

Authors: Mahendra Adhi Nugroho, Umar Yeni Suyanto, Didik Hariyanto, Septiningdyah Arianisari

PAGE 401 – 410

Paper 42: Comparative Review of Confidence and Other Evaluation Metrics in Predictive Modeling for Procurement Fraud Coalition

Authors: Saifuddin Mohd, Mohamad Taha Ijab

PAGE 411 – 418

Paper 43: Correcting Blue-Shift in Single-Image Dehazing via Haze-Compensated Von Kries Adaptation

Authors: Asniyani Nur Haidar Abdullah, Mohd Shafry Mohd Rahim, Sim Hiew Moi, Azah Kamilah Draman, Ahmad Hoirul Basori, Novanto Yudistira

PAGE 419 – 425

Paper 44: Enhancing Dermatological Diagnostics: An Enhanced Approach for Skin Cancer Classification Using pix2pix GAN

Authors: Adnan Afroz, Shaheena Noor, Shakil Ahmed Bashir, Umair Jilani

PAGE 426 – 433

Paper 45: A Hybrid AI Framework for DDoS Detection and Mitigation in SDN Environments Using CNN, GAN, and Semi-Supervised Learning

Authors: Abdelhakim HADJI, Brahim RAOUYANE

PAGE 434 – 439

Paper 46: Benchmarking Deep Learning Models for Visual Classification and Segmentation of Horticultural Commodities
Authors: Fuzy Yustika Manik, Syahril Efendi, Jos Timanta Tarigan, Maya Silvi Lydia
PAGE 440 – 448

Paper 47: Towards Designing a Blockchain-Based Model for E-Book Publishing
Authors: Maznun Arifa Mohammadan Makhtar, Novia Admodisastro, Suleymenova Laura Askarbekkyzy
PAGE 449 – 460

Paper 48: Adaptive Hybrid Deep Learning with Recursive Feature Elimination for Physical Violence Detection
Authors: Sukmawati Anggraeni Putri, Duwi Cahya Putri Buani, Achmad Rifa'i, Imam Nawawi
PAGE 461 – 468

Paper 49: Transformative Integration of Machine Learning in Software Applications in Light of Current Software Engineering Practices
Authors: Fawzi Abdulaziz Albalooshi
PAGE 469 – 479

Paper 50: Bridging Machine-Readable Code of Regulations and its Application on Generative AI: A Survey
Authors: Samira Yeasmin, Bader Alshemaimri
PAGE 480 – 490

Paper 51: Energy Efficient Workflow Allocation in Cloud Computing Using Improved Grey Wolf Optimization
Authors: Md. Mazhar Nezami, Anoop Kumar
PAGE 491 – 498

Paper 52: Feature Pyramid Network with Dual-Decoder Supervision for Accurate Stroke Lesion Localization in Multi-Modal Brain MRI
Authors: Satmyrza Mamikov, Zhansaya Yakhiya, Bauyrzhan Omarov, Yernar Mamashov, Akbayan Aliyeva, Balzhan Tursynbek
PAGE 499 – 509

Paper 53: Leveraging AI and Hybrid Intelligence for Robust Geospatial Data Fusion in Autonomous Terrestrial Navigation
Authors: Manel Salhi, Mounir Bouzguenda, Faouzi Benzarti, Fawaz Alanazi, Ezzeddine Touti
PAGE 510 – 520

Paper 54: SEARCHX: An Integrated Framework of Distributed Intelligent Search Services Based on Web Browser
Authors: Zehui Zhang, Lin Zhou, Jie Peng, Liwei Wang, Bo Cheng
PAGE 521 – 529

Paper 55: Multi-Criteria Using Dijkstra's Algorithm to Determine Optimal Time Paths in Vehicle Route Optimization
Authors: Basorudin, Handaru Jati, Nurkhamid, Puput Dani Prasetyo Adi
PAGE 530 – 538

Paper 56: Embedding Models: A Comprehensive Review with Task-Oriented Assessment
Authors: Lahbib Ajallouda, Meriem Hassani Saissi, Ahmed Zellou
PAGE 539 – 548

Paper 57: Machine Learning-Driven Emotional Feedback Analysis and Adaptive Content Generation for VR Movie and TV Users
Authors: Yun TANG
PAGE 549 – 560

Paper 58: Meta-Learning Prediction Framework for Asphalt Mixtures Fatigue Life Modeling
Authors: Longmeng Tan, Krzysztof Kowalski
PAGE 561 – 570

Paper 59: Quality Classification of Harumanis Mango Based on External Multi-Parameter and Machine Learning Techniques
Authors: Mohd Nazri Abu Bakar, Abu Hassan Abdullah, Muhamad Imran Ahmad, Norasmadi Abdul Rahim, Haniza Yazid, Wan Mohd Faizal Wan Nik, Shafie Omar, Shahrul Fazly Man@Sulaiman, Tan Shie Chow, Fahmy Rinanda Saputri
PAGE 571 – 578

Paper 60: Fourier Transform and Attention Guided Deep Neural Network for Face Anti-Spoofing in Medical Applications
Authors: Zhanseri Ikram
PAGE 579 – 588

Paper 61: A Technique for Automated Parallel Optimization of Function Calls in C++ Code
Authors: Shuruq Abed Alsaedi, Fathy Elbouraey Eassa, Amal Abdullah AlMansour, Lama Abdulaziz Al Khuzayem, Rsha Talal Mirza
PAGE 589 – 600

Paper 62: Unveiling the Drivers of Consumer Purchase Intention in Short-Form Video Marketing
Authors: Merisa Syafrina, Viany Utami Tjhin
PAGE 601 – 611

Paper 63: Integrative Hybrid Metaheuristic Algorithm for Hyperparameter Optimisation in Pre-Trained Convolutional Neural Network Models (I-HAHO)
Authors: Nazleeni Samiha Haron, Jafreezal Jaafar, Izzatdin Abdul Aziz, Mohd Hilmi Hasan, Muhammad Hamza Azam
PAGE 612 – 619

Paper 64: A Novel Performance-Based Time Series Forecast Combination Method and Applications with Neural Networks
Authors: M. Burak Erturan
PAGE 620 – 628

Paper 65: Integration of Color QR-Code Technology in Biometric Data Encoding and Facial Identity Systems
Authors: Nazym Kaziyeva, Kalybek Maulenov, Ruslan Ospanov, Abzhan Khamza Mukhtaruly
PAGE 629 – 637

Paper 66: Enhancing the Scanability of Damaged QR Codes Through Image Restoration Using GANs Combined with the Spectral Normalization Technique
Authors: Puwadol Sirikongtham, Apichaya Nimkoompai
PAGE 638 – 644

Paper 67: Recent Integrating Machine Learning and Malay-Arabic Lexical Mapping for Halal Food Classification
Authors: Noorrezam Yusop, Massila Kamalrudin, Nuridawati Mustafa, Tao Hai, Mohd Nazrien Zaraini, Halimaton Hakimi, Siti Fairuz Nurr Sardikan
PAGE 645 – 652

Paper 68: Evaluating Head Pose Estimation for Assessing Visual Attention in Children with Special Needs During Robot-Assisted Therapy
Authors: Rusnani Yahya, Rozita Jailani, Nur Khalidah Zakaria, Fazah Akhtar Hanapiah
PAGE 653 – 662

Paper 69: A Novel Taxonomy for Human Activity Recognition Based on a Systematic Analysis of Public UAV Datasets
Authors: Sumaya Abdulrahman Altuwairqi, Salma Kammoun Jarraya
PAGE 663 – 687

Paper 70: Evaluation of the Impact of Cybersecurity Knowledge on the Prevention of Social Cybercrime Among University Students in Mexico, Colombia, and Peru
Authors: Yasmina Riega-Viru, Lainiver Mendoza Munar, Mario Ninaquispe-Soto, Kiara Nilupu-Moreno, Juan Luis Salas-Riega, Alfonso Renato Vargas-Murillo, Yolanda Pinto Bouroncle
PAGE 688 – 695

Paper 71: Understanding Echo Chambers in Recommender Systems: A Systematic Review
Authors: Meriem HASSANI SAISSI, Nouhaila IDRISSI, Ahmed ZELLOU
PAGE 696 – 711

Paper 72: Critical Review of Object Detection Techniques for Traffic Light Detection in Intelligent Transportation Systems
Authors: Adhwa Salemi, Muhammad Arif Mohamad
PAGE 712 – 717

Paper 73: Conversational AI-Powered VR Development Model for Tourism Promotion in Thailand: Expert Assessment and Stakeholder Acceptance
Authors: Jenasama Srihirun, Kridsanapong Lertbumroongchai, Vitsanu Nittayathammakul, Pimon Kaewdang
PAGE 718 – 728

Paper 74: Roadmap for Emerging Cyberbullying Mitigation: Integrating AI-Based Solutions, Ethics, and Policy
Authors: Atif Mahmood, Shaik Shabana Anjum, Umm E Mariya Shah, Pavani Cherukuru, Javid Iqbal, Sarah Bukhari
PAGE 729 – 739

Paper 75: Intelligent Visualization and Knowledge Graph Analysis for Trend Detection
Authors: Sunan Lv
PAGE 740 – 750

Paper 76: An Efficient and Scalable Reinforcement Learning-Driven Intelligent Resource Management and Secure Framework for LoRaWAN
Authors: Shaista Tarannum, Usha S. M
PAGE 751 – 761

Paper 77: Real-Time Multi-Scale Object Detection in Surveillance Using Hybrid Transformer Architecture
Authors: Roshan D Suvaris, Rahul Suryodai, S. Narayanasamy, Aanandha Saravanan, Raman Kumar, P N V Syamala Rao M, Elangovan Muniyandy
PAGE 762 – 774

Paper 78: Aerial Draft Surveyor (ADS)
Authors: John Matthew H. Escarro, Fharjan M. Taguinopon, Gyrielle Kysha M. Demegillo, Dan Kevin T. Amper, Rosanna C. Ucat, Mark John S. Pag-Alaman
PAGE 775 – 785

Paper 79: Privacy-Aware Federated Graph Neural Networks for Adaptive and Explainable Cancer Drug Personalization
Authors: Tripti Sharma, Lakshmi K, M. Misba, Jasgurpreet Singh Chohan, R. Aroul Canessane, Komatigunta Nagaraju, Adlin Sheeba
PAGE 786 – 798

Paper 80: Hybrid Vision Transformer and MLP-Mixer for Epileptic Seizure Detection in Intracranial EEG
Authors: Thouraya Guesmi, Abir Hadriche, Nawel Jmail
PAGE 799 – 807

Paper 81: Truth Under Pressure: A Deep Learning-Based Lie Detection System for Online Lending Using Voice Stress and Response Latency
Authors: Ahmad Ihsan Farhani, Alhadi Bustamam, Rinaldi Anwar, Titin Siswantining
PAGE 808 – 817

Paper 82: Classification of Mangrove Ecosystem Health Using Sentinel-2 Images with Genetic Algorithm Optimization in Machine Learning Algorithms
Authors: Putri Yuli Utami, Murni Ramadhani, Rudi Alfian, Barry Ceasar Octariadi, Dimas Kurniawan
PAGE 818 – 835

Paper 83: User Identity Confirmation Property Management System Based on State Secret Algorithm and Blockchain Technology
Authors: Xiao Tian, Xing Chen
PAGE 836 – 848

Paper 84: Coordination, Communication and Robustness in Multi-Agents: An Industrial Network Scenario Using Trust Region Policy Optimization
Authors: Munam Ali Shah
PAGE 849 – 860

Paper 85: Spatiotemporal Graph Networks for Relational Reasoning in Campus Infrastructure Management
Authors: Sanjay Agal, Krishna Raulji, Nikunj Bhavsar, Pooja Bhatt
PAGE 861 – 872

Paper 86: A Voting-Based Ensemble Method for Deep Learning Performance Enhancement
Authors: Mohammed Abdel Razek, Rania Salah El-Sayed, Arwa Mashat, Shereen A. El-aal
PAGE 873 – 885

Paper 87: A Spatiotemporal Forex Trading System Based on a Hybrid Model GAT-LSTM: Forecasting Forex Price Directions
Authors: Nabil MABROUK, Marouane CHIHAB, Younes CHIHAB
PAGE 886 – 895

Paper 88: Experimental Validation of an Adaptive Controller for a Mecanum-Wheel Robot with Unknown Center-of-Gravity Offset and Slope Inclination
Authors: Chawannat Chaichumporn, Supaluk Prapan, Nghia Thi Mai, Md Abdus Samad Kamal, Iwanori Murakami, Kou Yamada
PAGE 896 – 905

Paper 89: Enhancing Out-of-Distribution Detection for Retail Time-Series Data Using Entropic Methods
Authors: Nga Nguyen Thi, Tuan Vu Minh, Khanh Nguyen-Trong
PAGE 906 – 914

Paper 90: A Review of Artificial Intelligence in Inventory Management: Methods, Applications and Directions
Authors: Jinjin Li, Huijun Huang, Yuping Gong, Lei Wang, Xiangui Yin, Yichang Liu
PAGE 915 – 928

Paper 91: Glioma Classification Using Harris Hawks-Driven Optimized Gradient Boosting Classifier Along with SHAP-Based Interpretability
Authors: SM Naim, Jun-Jiat Tiang, Abdullah-Al Nahid
PAGE 929 – 939

Paper 92: Federated Performance-based Averaging (FedPA): A Robust and Selective Learning Framework for Chest X-Ray Classification in Heterogeneous Data Environments
Authors: Atif Mahmood, Tashin Khan Sadique, Saaidal Razalli Azzuhri, Roziana Ramli, Leila Ismail
PAGE 940 – 952

Paper 93: Privacy-Preserving Education Data Sharing Scheme Based on Consortium Blockchain
Authors: Jiaqi Guo, Zhuoran Wang, Ningning Liu
PAGE 953 – 963

Paper 94: RT-DETR Edge Deployment: Real-Time Detection Transformer for Distracted Driving Detection
Authors: Fares Hamad Aljahani
PAGE 964 – 975

Paper 95: Breast Cancer Classification Using Ensemble Voting: A Feature Selection Approach
Authors: Antu Kumar Guha, Jun-Jiat Tiang, Abdullah-Al Nahid
PAGE 976 – 984

Paper 96: A Two-Stage Framework for Abnormalities Detection in WCE Images by Combining Semantic Segmentation and Deformable Agent-Based Classification
Authors: Brahim Alibouch, Yasmina El Khalfaoui
PAGE 985 – 992

Paper 97: Predicting Employee Attrition in the Saudi Private Sector Using Machine Learning
Authors: Haya Alqahtani, Hana Almagrabi, Amal Alharbi
PAGE 993 – 1002

Paper 98: Optimizing Asset Transfer Process in ERP Using Business Process Management Technique
Authors: Ravindu Yasarathne, Naduni Ranatunga, Vikasitha Herath, Lakshan Chalinda, Chathurangika Kahandawaarachchi, Sanjeeva Perera, Chamath Randula
PAGE 1003 – 1023

Paper 99: Stochastic Policies, Deterministic Minds: A Calibrated Evaluation Protocol and Diagnostics for Deep Reinforcement Learning
Authors: Sooyoung Jang, Seungho Yang, Changbeom Choi
PAGE 1024 – 1034

Paper 100: Expert Systems in Tuberculosis Prevention Established in Certainty Factor
Authors: Inooc Rubio Paucar, Cesar Yactayo-Arias, Laberiano Andrade-Arenas
PAGE 1035 – 1047

Paper 101: Feature-Optimized Machine Learning for High-Accuracy Ammunition Detection in X-Ray Security Screening
Authors: Osama Dorgham, Nijad Al-Najdawi, Mohammad H. Ryalat, Sara Tedmori, Sanad Aburass
PAGE 1048 – 1059

Paper 102: Evaluating Transparency in the Development of Artificial Intelligence Systems: A Systematic Literature Review
Authors: Giulia Karanxha, Paulinus Ofem
PAGE 1060 – 1072

Paper 103: Auditable Real-Time Cold-Chain Monitoring with IoT and Blockchain Anchoring
Authors: Mohamed DOUBIZ, Mouad BANANE, Abdelali ZAKRANI, Allae ERRAISSI
PAGE 1073 – 1082

Paper 104: Systematic Literature Review of Reactive Jamming Attacks Mitigation Techniques in Internet of Things Networks
Authors: Enos Letsoalo, Topside Mathonsi, Tshimangadzo Tshilongamulenzhe, Daniel du Duplesis
PAGE 1083 – 1090

Paper 105: Virtual Assistant Based on Recurrent Neural Networks: Scope and Coverage of Health Insurance
Authors: Diego Alberto Paz-Medina, David Eduardo Rojas-Cavassa, Ernesto Adolfo Carrera-Salas
PAGE 1091 – 1104

Paper 106: Text Information Data Mining Method in Natural Language Processing Tasks
Authors: Shengguo Guo, Dandan Xing
PAGE 1105 – 1114

Paper 107: Document Similarity Detection for Project Development Using Fused Interactive Attention Mechanisms
Authors: Chao Zhang, Ying Zhang, Gang Yang, Fan Hu
PAGE 1115 – 1124

Paper 108: A Comparison Between FAHP-TOPSIS and FAHP-FTOPSIS Methods for Selecting the Best Products for Home-Based Sellers: A Performance Analysis
Authors: Selvia Lorena Br Ginting, Zulaiha Ali Othman
PAGE 1125 – 1135

AI-Assisted Workflow Optimization and Automation in the Compliance Technology Field

Zhen Zhong
Senior Data Engineer, Graduate School of Arts & Sciences, Georgetown University, Washington, DC 20001, United States

***Abstract*—Against the backdrop of digital transformation and stricter regulation, enterprise compliance work demands higher efficiency and accuracy. The auxiliary compliance process has become an important entry point for optimizing the compliance system due to its strong transactional nature and high degree of repetition. This study focuses on the process characteristics of auxiliary compliance work, sorts out its structural composition and organizational mechanism, proposes an optimization path with process reengineering, system modeling, and technology integration as the core, and focuses on exploring the collaborative application of key technologies such as RPA, rule engine, and semantic recognition in process automation. Research suggests that the systematic optimization and intelligent upgrading of auxiliary processes will help build a modern compliance operation system that is responsive, efficient, structurally clear, and risk controllable.**

Keywords—*Compliance technology; auxiliary process; process optimization; automation*

I. Introduction

With the continuous escalation of compliance requirements worldwide, multinational corporations face growing complexity in aligning operations with diverse regional regulations. Existing studies have applied robotic process automation (RPA), rule engines, and semantic recognition to improve efficiency in transactional compliance, particularly within finance and healthcare sectors in Europe and North America. However, these approaches remain limited by dependence on structured data, poor adaptability to rapidly changing regulations, and weak cross-industry transferability.

Most research focuses on static rule-based automation, while practical compliance increasingly demands dynamic adjustment to new laws and contextual anomalies. Organizations, therefore, struggle to maintain both precision and agility. Future advancements should integrate AI-driven anomaly detection, real-time regulatory monitoring, and adaptive policy learning. Expanding these capabilities beyond finance and healthcare to manufacturing, logistics, and telecommunications could enhance global compliance resilience. Auxiliary compliance processes, with their high frequency and standardization, offer a strategic foundation for achieving this integration. This study introduces an integrated compliance automation framework that combines adaptive threshold control, RPA orchestration, and semantic rule modeling—an approach not yet systematically implemented in prior literature.

II. Process Characteristics of Auxiliary Compliance Work

A. Logical Division of Task Types and Execution Methods

In international corporate compliance management, auxiliary compliance tasks are generally categorized into data collection and entry, document verification and archiving, standardized report generation, rule matching, and anomaly flagging. These tasks typically rely on clearly defined rules and fixed-step operational patterns. For example, in anti-money laundering (AML) compliance within the financial sector, the backend system may pull data from multiple sources to match customer identity information, followed by compliance list screening carried out by either human operators or automated tools. In the medical data processing domain, compliance teams perform formatting and privacy redaction of electronic medical records in accordance with regulations such as HIPAA. These tasks exhibit a high degree of divisibility and repeatability, making them well-suited for partial or complete automation to reduce human error and accelerate processing [1].

B. Dimensions for Identifying Typical Processes and Operational Bottlenecks

The typicality of auxiliary compliance processes can be identified through dimensions such as task frequency, data structure complexity, cross-departmental interaction volume, and the rate of regulatory change. In international cross-border payment compliance reviews, high-frequency transaction screening and cross-system data matching often become bottlenecks due to inconsistent data formats and lengthy information transmission chains. In the energy sector's carbon emissions compliance audits, bottlenecks are concentrated in the parsing and standardization of unstructured report texts. By identifying these dimensions, it becomes possible to pinpoint the process segments most amenable to technological intervention for efficiency improvement, thereby providing targeted guidance for subsequent process reengineering and automation deployment.

C. Process Organization Characteristics of Manual Processing and Data Structure

Under traditional models, auxiliary compliance processes often rely on human operators for task scheduling and result confirmation, with data handling structures typically characterized by decentralization and non-real-time processing. For example, in international insurance claim compliance reviews, staff must switch between multiple business systems to obtain complete information, resulting in extended task completion times and difficulties in ensuring data consistency.

In the telecommunications sector's user privacy compliance management, manual workflows frequently involve handling semi-structured data, such as customer communication logs and transaction receipts, which increases data cleansing costs and limits real-time risk response capabilities. These organizational characteristics underscore the necessity of strengthening data structure standardization and system interoperability in process design, laying the groundwork for subsequent automated execution.

III. Systematic Design for Assisting Workflow Optimization

A. Underlying Logic of Task Decomposition and Process Reengineering

In cross-border KYC/AML scenarios, auxiliary compliance workflows can be abstracted as a directed acyclic graph G=(V,E)，where each task v∈V is decomposed into atomic operations $v \rightarrow \{a_k\}$ and designated for either human (H) or robotic (R) execution, with $x_v \in \{H,R\}$. The optimization objective is defined as [Formula (1)]:

$$\min J = \alpha \cdot T(G,x) + \beta \cdot \varepsilon(G,x) + \gamma \cdot C(G,x) \quad (1)$$

subject to [Formula (2)]:

$$s(u) + t_u \le s(v)(u,v) \in E、\ \tau(v) \le \hat{\tau}(SLA) \quad (2)$$

and compliance predicate [Formula (3)]:

$$\Phi(d) = true(GDPR / HIPAA) \quad (3)$$

Incoming data are standardized through a mapping function (d)=σ(φ(d)) before entering the pipeline, ensuring that privacy processing $\pi(d) = mask(d,k)$ and kkk-anonymity (k≥k0). requirements are met [2].

A rule engine R={r_i} drives decision-making, where:

$$r_i : p_i(d) \ge \theta \Rightarrow action_{...} \quad (4)$$

and the probability p_i is computed by a semantic model. The scheduler allocates tasks between RPA and human reviewers based on critical path analysis and queue load: when $p_i<\theta$ or the incremental risk Δrisk(v)>δ, x_v is switched to H and a secondary audit is triggered. In the objective function, T(G,x) represents the total task completion time determined by workflow structure G and execution mode x; ε(G,x) denotes the cumulative error rate from both automated and manual tasks; C(G,x) is the overall operational cost, including human labor and system resource usage. Coefficients α,β,γ define their relative weighting according to compliance priorities—speed, accuracy, or cost efficiency. The constraint s(u)+t u≤s(v) ensures task sequencing consistency, preventing logical overlap in dependent operations. τ(v)≤τ^enforces service-level compliance for each task, while the predicate Φ(d)=true guarantees that all data handling steps conform to privacy laws such as GDPR and HIPAA. The parameter θ is a dynamic confidence threshold determining whether decisions are automated or escalated, and Δrisk(v) quantifies deviations in risk estimates used to trigger human intervention. This mathematical formalization provides a unified framework for optimizing efficiency, accuracy, and compliance integrity across heterogeneous processes. For example, in a European retail bank's account opening process, the sequence of passport OCR → entity parsing → sanctions list screening → risk stratification can be reengineered under this model, resulting in significant reductions in both processing time T and error rate ε, while meeting SLA requirements. Unlike prior work that isolates automation from compliance logic, this framework unifies workflow optimization, real-time feedback control, and rule-based governance into a single adaptive architecture.

B. Process Construction Path Based on Scenario Logic and Modeling Rules

Following the completion of task decomposition and process reengineering logic design, the construction of scenario logic and modeling rules becomes the key step in transforming the abstract model into an executable workflow. This process involves mapping the atomized task units and execution nodes to concrete business scenarios, ensuring that the workflow achieves automation and efficiency while meeting regulatory requirements [3]. For example, in the compliance review process for account opening at a UK retail bank, the event of a customer submitting an identity document triggers the system to call the OCR module for image parsing. The parsed result is immediately sent to the sanctions list screening engine, which applies predefined rules to determine the presence of high-risk matches. When the match confidence falls below 0.90, the workflow automatically switches to the manual review path and simultaneously generates a complete processing record in the backend to ensure subsequent compliance with regulatory traceability requirements.

The formulation of modeling rules must balance the mandatory nature of legal provisions with the operational flexibility of the institution, typically comprising three categories: legal provisions, internal control rules, and model thresholds, all unified into executable policies. For instance, when processing electronic medical records, a large U.S. healthcare group embeds the HIPAA-mandated PHI field list into the rule engine as the legal layer rules, ensuring that all identified sensitive fields are automatically redacted before data transmission. At the internal control layer, fields deemed highly sensitive must undergo verification through a dual-review mechanism; at the model threshold layer, the system requires the named entity recognition model to achieve a confidence score of no less than 0.92 before automatically masking the identified entity. All rules must pass static conflict detection and priority sequencing prior to deployment to prevent execution conflicts. Through unified interface calls, the workflow can operate stably across systems, with historical case playback and canary testing used to validate the feasibility and accuracy of the strategy. This approach establishes a compliance workflow system that is scalable, interpretable, and capable of rapid iteration.

C. Process Collaboration Mechanism and Feedback Loop Design

The collaboration and feedback loop are driven by observable metrics to enable adaptive adjustment of thresholds and resources. The core approach is to continuously monitor key runtime indicators—end-to-end latency T, false positive rate (FPR), false negative rate (FNR), and backlog volume B—and

to consolidate them into a single weighted-deviation objective for parameter tuning [Formula (5)]:

$$(J = \sum w_k (m_k - m_k^{\backslash *})^2) \tag{5}$$

Node load can be controlled with a single constraint:

$$\rho = \lambda / (c\mu) \le \rho_{\max} \tag{6}$$

where, λ is the arrival rate, μ is the service rate, and c is the number of concurrent channels (including both RPA bots and human reviewers). Routing uses a risk score threshold θ: when s≤θ, processing is automated; when s>θ, the case is routed to manual review. Both θ and the concurrency c of each node are incrementally adjusted by the controller following the principle of "the greater the deviation, the stronger the adjustment," while remaining within feasibility and compliance constraints such as SLA adherence, audit logging, and data minimization. For example, in the KYC account opening process at a UK retail bank, the sanctions list screening node recorded a load of ρ=0.86 during peak hours, exceeding the alert threshold of 0.80 [4]. The system first increased concurrency for that node from 8 to 12 and temporarily reduced OCR node concurrency to maintain the overall capacity cap. Within the following 10 minutes, the FPR rose to 3.1 per cent, prompting the controller to fine-tune θ from 0.88 to 0.84, increasing the manual review proportion to 22 per cent, bringing the false positive rate back within the 2.0 per cent target range and clearing the backlog. Data distribution drift is monitored using the Population Stability Index (PSI). When the upper PSI threshold is triggered, the system switches to "conservative mode", freezing θ adaptation, allowing only capacity-side fine-tuning, and increasing the sampling rate for high-confidence cases. All routing and parameter changes are simultaneously recorded in the audit log (sample ID, rule version, evidence chain) to ensure full traceability and reproducibility.

IV. Key Technology Path for Compliance Process Automation

A. Task Substitution and Process Execution Based on RPA

Building on the above foundations of task atomization and scenario–rule executable design, RPA functions as the "executor", transforming abstract nodes into controllable automated actions and integrating feedback loop parameters (thresholds and concurrency) to form an adaptive operating layer. The implementation path uses BPMN/DMN as the primary framework, registering task modules such as "document parsing, list screening, risk categorization, and audit archiving" into the orchestrator, with clearly defined input/output contracts and error semantics. Each task module is linked to a pool of unattended bots and manual review seats, with the scheduler dynamically allocating concurrency c according to the node load Formula (6), and using the routing threshold θ to control the proportion of automated versus manual processing (risk score s≤θ proceeds automatically, while scores above the threshold are routed to manual review).

Before data enters the RPA, minimization and masking are applied. Idempotent keys (sample_id + hash) ensure "exactly-once" execution, with failed executions applying exponential backoff and entering a dead-letter queue for manual takeover. Cross-system calls use short-lived tokens and a secure key vault to limit permission scope and access duration. To support iterative updates of rules and models, all bots are packaged and versioned as container images, with sandbox replays of historical cases comparing three key indicators—processing latency, false positive/false negative rates, and backlog volume—before gradual rollout through gray traffic. During runtime, telemetry continuously reports node latency, queue lengths, and error codes, enabling the controller to fine-tune θ and c, while writing parameters and evidence chains into an immutable audit log. For example, in the KYC account opening process of a UK retail bank, the "passport upload" event triggers an unattended RPA to call an OCR service, outputting document layout structures and entity candidates. These entities are matched against sanctions lists in the rule engine; if the list similarity score is 0.86 and below the threshold θ=0.90, the case follows the automated approval branch and generates an audit record. During peak hours, when the sanctions list screening node load ρ=0.83 approaches its limit, the orchestrator increases concurrency for that node from 8 to 12 and temporarily reduces OCR node concurrency to maintain the overall capacity cap, bringing end-to-end latency back within 180s in 10 minutes. If the false positive rate briefly rises above 3 per cent, the controller fine-tunes θ to 0.84, increasing the manual review share until the FPR returns to target range, after which the parameters automatically revert.

In the exception path, anti-crawling and document tamper checks are performed by a security-aware bot before processing, triggering high-risk flags and secondary review if needed. Cross-border address verification is handled via a geosanction API call by the bot, which retrieves real-time lists and caches signed snapshots to meet traceability requirements. To enhance explainability, each automated approval or rejection outputs a "data version – rule version – model threshold – evidence snippet" tuple for audit sampling and regulatory inquiries. Fig. 1, UK Retail Bank KYC Automation Orchestration Diagram (Event → Task Module → Routing → Feedback Loop Adjustment) may be cited to illustrate the interactions between the orchestrator, bot pool, manual stations, and monitoring/logging components.

B. Mechanism for Structured Conversion of Text Data

For unstructured text in cross-border compliance contexts (such as scanned documents, PDFs, emails, and chat logs), the conversion mechanism is implemented as a traceable pipeline consisting of "ingestion – parsing – extraction – alignment – validation – storage", integrated with the previously described threshold and concurrency control. In the ingestion stage, fingerprint deduplication and language detection are performed, generating idempotent keys and page coordinate indexes, with sensitive segments undergoing minimization and partial masking before processing. In the parsing stage, layout analysis is combined with adaptive OCR to differentiate between paragraphs, tables, and key-value areas, while a document-type classifier routes passports, bank statements, medical records, and other documents to the corresponding template families or template-free extractors.

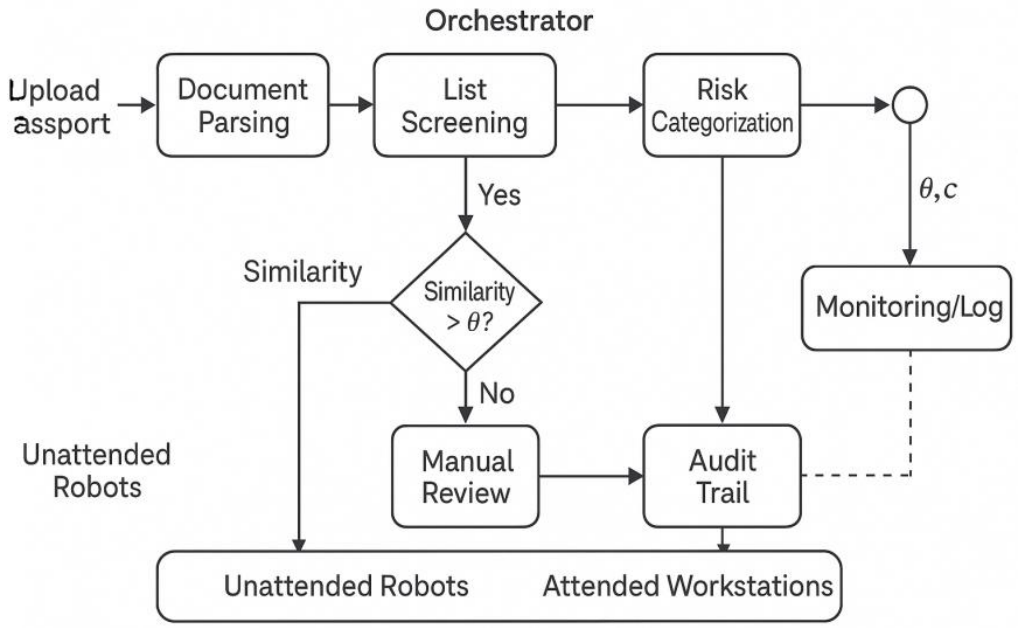


Fig. 1. UK retail bank KYC automation orchestration diagram.

The extraction stage operates under a hybrid "rule + model" paradigm: regular expressions and dictionaries ensure high-precision identification of strongly formatted fields such as ID numbers, IBAN, SWIFT codes, and CPT/ICD codes, while named entity recognition and relationship extraction models cover names, addresses, organizations, dates, and transactional semantics. These outputs include field-level confidence scores and evidence fragments (page number, bounding box, matching path). In the alignment stage, candidate entities are mapped to a unified business schema and external standards (ISO-8601 dates, ISO-3166 country/region codes, ISO-4217 currency codes, LEI/SSN placeholders), with cross-document merging achieved through entity resolution. Blocking keys and similarity metrics within the threshold range trigger manual review [5].

During validation, both field-level and sample-level thresholds are applied for consistency checks and compliance predicate evaluations (data minimization, retention periods, cross-border transfer flows). Exceptions are routed to compensating transactions and dead-letter queues while generating an auditable evidence chain. In the storage stage, versioned JSON Schemas and event sourcing are used to record evidence packages, rule versions, and model versions, enabling playback and incremental deployment.

For example, in the de-identification process of electronic health records (EHR) at a U.S. hospital group, the system ingests mixed HL7/PDF records, uses layout analysis to locate tables and free-text areas, and employs PHI detection models in combination with dictionary rules to output candidates for names, addresses, phone numbers, and geographic identifiers along with confidence scores. When an address confidence score falls below the set threshold, the case is routed to a manual workstation for verification, while high-confidence fields are directly masked or replaced with pseudo-random values. All changes are logged in an immutable audit record containing the sample ID, field, evidence fragment, and audit signature. Field-level false positive and false negative statistics are fed back to the monitoring dashboard to adjust sampling rates and thresholds online, ensuring high-quality, interpretable structured output under GDPR/HIPAA constraints [6].

C. Integrated Design of Rule Engine and Automated Decision-Making Mechanism

The integration approach uses a three-layer rule stack—"legal hard constraints – internal policies – model scoring"—to drive decisions. DMN/decision tables and executable DSL are employed to define conditions, actions, and priorities. At runtime, external lists, thresholds, and black/white lists are mounted as snapshot versions to ensure replayability and traceability. The decision process consists of pre-check, rule matching, conflict resolution, and action orchestration: the pre-check layer verifies data completeness and timeliness; the matching layer triggers based on a combination of static rules and dynamic thresholds; the conflict resolution layer applies a priority order of "regulatory rejection > risk block > business approval"; and the action layer works with the orchestrator to issue approvals, rejections, or "enhanced verification" directives [7].

The risk score s output from the model and the current operating threshold θt act as soft evidence in the rules, while any "hard" regulatory matches override the model's decision. Change management uses policy versioning and case replay, with all new rules required to pass unit test suites, historical sample shadow runs, and gray (canary) releases. Online monitoring of FPR/FNR, average decision latency, and trigger rates ensures stability.

For example, in a PSD2/SCA and AML scenario at a European payment institution, the rule table specifies that "originating from a high-risk country + abnormal device fingerprint + transaction amount exceeding the SEPA threshold" triggers "enhanced identity verification". If a sanctions list hit

occurs simultaneously, the transaction is rejected outright. When the daily false positive rate increases, operations adjust only θt and the sanctions list similarity threshold without altering the legal layer rules. The decision explanation is returned to the audit system in the format "active rule ID – evidence fragment – version number", ensuring interpretability, controllability, and iterative capability [8].

V. Conclusion

Against the backdrop of increasingly stringent global compliance requirements, building an automation system centered on process reengineering, scenario logic modeling, and key technology integration can simultaneously improve execution efficiency and strengthen risk control capabilities. RPA-based task substitution, structured conversion of text data, and the rule engine's automated decision-making mechanism—supported by feedback loops and collaborative scheduling—enable balanced node load, adaptive thresholds, and full traceability. Validated through cases in UK retail banking and healthcare institutions in Europe and North America, the system has demonstrated strong stability and scalability in controlling processing latency, false positives, and false negatives, offering a replicable technical pathway for cross-industry and cross-regional compliance operations.

References

[1] Hasanpour A M, Kirkegaard M, Fafoutis X. EdgeMark: An automation and benchmarking system for embedded artificial intelligence tools[J].Journal of Systems Architecture,2025,167103488-103488.

[2] Herold A, Lo C W, Sharp A, et al. Enhancing MRI efficiency in musculoskeletal examinations: Impact of optimized facility design and workflow optimization efforts[J].European Journal of Radiology,2025,190112263-112263.

[3] Smith J, Fishman K E, Chu C L, et al. From Automation to Innovation: How AI is Reshaping Global Industries.[J]. Journal of the American College of Radiology: JACR,2025.

[4] Mahalingam T, S. N, T. S, et al. AI Assisted Studying Partner for Students[J].International Journal of Advanced Research (IJAR),2024,12(3).

[5] Adil K, A.V.V.S. S, Alaa K, et al. Wireless communication based cloud network architecture using AI assisted with IoT for FinTech application[J].Optik,2022,269.

[6] LCPtracker Inc.; LCPtracker Introduces Advanced Mobile Technology to Align Field-to-Office Compliance Reporting[J].Journal of Engineering,2020,892-.

[7] Dynamic Comply Launches AI Governance Solutions for ISO 42001, NIST AI RMF, EU AI Act, and other Compliance Frameworks[J].M2 Presswire,2025.

[8] Resilinc Debuts Agentic AI Platform for Supply Chain Risk and Compliance[J].Food and Beverage Close - Up,2025.

From Legacy to Cloud: Migration Strategies for Traditional Financial Institutions Using AWS

Uday Kiran Chilakalapalli[1], Brij Mohan[2], Vinodkumar Reddy Surasani[3]
Sr Analyst, LPL Financial, Fort Mill, SC, Department of Data Science and Analytics, Georgia State University, USA[1]
VP-Principal Software Dev, LPL Financial, Fort Mill, SC, USA[2]
Sr. Software Engineer, RBC Wealth Management, MN, USA[3]

***Abstract*—Traditional financial institutions face unprecedented pressure to modernize their technological infrastructure while maintaining regulatory compliance and operational stability. This research examines the strategic approaches, implementation challenges, and outcomes of migrating legacy banking systems to Amazon Web Services (AWS) cloud infrastructure through a mixed-methods analysis of twelve financial institutions that completed migrations between 2019 and 2024. Through structured interviews with technology leaders and quantitative analysis of migration outcomes, including regulatory considerations and real-world implementation cases, this study identifies key success factors and potential pitfalls in large-scale financial services cloud adoption. The research reveals that institutions adopting phased migration strategies with robust risk management frameworks achieve 92% success rates with 30-45% cost reductions and 40-60% performance improvements, compared to 58% success rates for rapid, wholesale transitions. Furthermore, the study demonstrates that AWS-specific services such as AWS Control Tower and AWS Config provide essential governance capabilities that traditional financial institutions require for regulatory compliance during cloud transformation initiatives.**

***Keywords*—Cloud migration; financial services; AWS; compliance; risk management; governance**

I. Introduction

The financial services industry stands at a technological crossroads. Decades of regulatory requirements have created complex, interconnected legacy systems that now serve as barriers to innovation rather than foundations for growth. Meanwhile, emerging fintech companies leverage cloud-native architectures to deliver superior customer experiences at reduced operational costs [8]. Traditional financial institutions must navigate this transformation carefully, balancing the imperative for modernization against the risks inherent in migrating mission-critical financial systems. Amazon Web Services has emerged as a leading platform for financial services transformation [1], offering specialized tools and compliance frameworks designed specifically for regulated industries [2]. However, the path from legacy mainframe systems to cloud-native architectures presents unique challenges that extend beyond technical considerations to encompass regulatory compliance [3], risk management [4], and organizational change.

Guiding research questions. How can traditional financial institutions successfully migrate their legacy systems to AWS while maintaining operational integrity and regulatory compliance? This question sits at the intersection of four dimensions developed throughout the paper: 1) migration architecture and patterns, 2) regulatory compliance and governance, 3) risk management and operational resilience, and 4) organizational change management.

Scope. We focus on regulated institutions (banks, credit unions, insurers) executing AWS migrations between 2019–2024 in North America and Europe, where supervisory guidance (e.g., FFIEC, BCBS, EBA, PRA) is most mature. Multi-cloud strategies are noted but not evaluated in depth; fintech-native startups are out of scope.

Sub-questions. 1) Which migration strategies (rehost/replatform/refactor) correlate with better outcomes under regulatory constraints? 2) Which AWS governance and security services (e.g., Control Tower, Config, GuardDuty) most effectively support compliance controls? 3) What risk-management practices reduce incidents during transition (e.g., data integrity checks, parallel run, DR drills)? 4) What business outcomes (cost, performance, resilience, time-to-market) are realized post-migration?

Hypothesis. Institutions that adopt phased migrations with formal governance (AWS Control Tower/Config) and continuous risk controls achieve superior outcomes versus big-bang transitions. We operationalize this question using the research methodology summarized in Fig. 1 and evaluate twelve institutions to derive evidence-based recommendations.

A. Research Objectives

The primary objectives of this research include analyzing the strategic approaches used by traditional financial institutions for AWS migration, identifying critical success factors and common failure modes in financial services cloud transformation, examining the role of AWS-specific services in addressing regulatory and compliance requirements, developing a framework for risk-managed migration strategies tailored to financial institutions, and evaluating the long-term business impact of cloud transformation on traditional financial services. As illustrated in Fig. 1, the research methodology framework analyzes four key dimensions critical to financial services cloud migration success: technical architecture, regulatory compliance, risk management, and organizational factors.

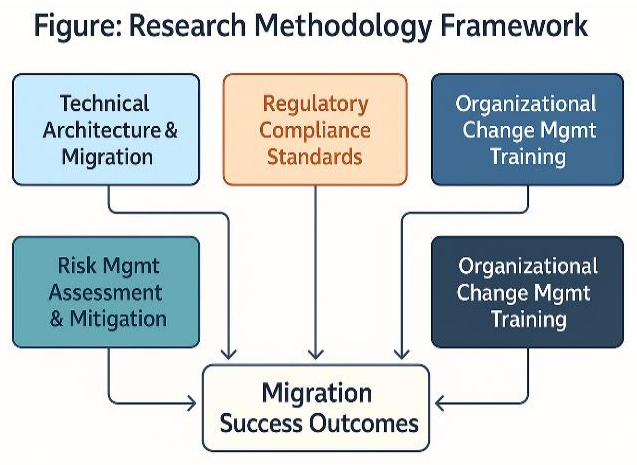


Fig. 1. Research methodology framework.

The remainder of this study is organized as follows: Section II reviews the relevant literature on cloud adoption in financial services and legacy system modernization. Section III describes the mixed-methods research approach and data collection procedures. Section IV examines AWS services architecture for financial institutions, while Section V analyzes migration strategies and risk management approaches. Section VI addresses regulatory compliance frameworks. Section VII presents two detailed case studies. Sections VIII and IX cover technical implementation challenges and organizational change management, respectively. Section X presents results and analysis of migration outcomes, Section XI presents the performance analysis. Section XII discusses strategic implications, and Section XIII concludes with key findings and recommendations.

II. Literature Review

A. Cloud Adoption in Financial Services

The literature on cloud adoption within the financial services sector has evolved rapidly in recent years. Early studies emphasized security and regulatory risks as primary obstacles to adoption, with some scholars recommending that financial institutions retain on-premises infrastructure to maintain data control [5]. However, more recent research highlights a shift in institutional attitudes and regulatory guidance. For example, the U.S. Department of the Treasury [6] reported that cloud services now play a critical role in the modernization of financial institutions, though transparency, interoperability, and concentration risk remain key concerns. Institutions are increasingly recognizing cloud adoption as vital to achieving operational efficiency and digital innovation.

A 2023 systematic review by Javaheri et al. [7] examined cybersecurity threats in the FinTech domain and identified a growing reliance on cloud-native solutions, often accompanied by elevated risks in data access and compliance. This underscores the importance of strong governance mechanisms such as AWS Control Tower and AWS Config, which have emerged as pivotal tools to satisfy regulatory oversight requirements [1]. The literature has also begun to reflect a more nuanced understanding of the regulatory environment. Recent guidance from FFIEC and the European Banking Authority (EBA) encourages cloud adoption under stringent governance and monitoring protocols [3], [17].

The adoption of structured cloud governance frameworks has shown measurable benefits. According to the Financial Services Cloud Consortium's 2023 report, institutions with formal governance structures experienced 40% fewer security incidents during migration than those without them. These findings are reinforced by studies highlighting the importance of comprehensive migration planning and risk controls [6]. A recent case study by Johnson et al. [47] demonstrated how high-revenue financial institutions can optimize both cost and efficiency outcomes through a structured migration framework. Additionally, Jain and Singh [41] proposed a model to estimate and optimize cloud migration costs, offering quantifiable benefits for financial services undergoing transformation.

A 2024 survey by the Cloud Security Alliance [43] further reveals that financial institutions prioritizing cyber resiliency in cloud adoption reported significant improvements in governance compliance and incident response readiness.

B. Legacy System Modernization

Legacy system modernization remains one of the most technically and operationally complex challenges for financial institutions. Traditional banking infrastructure often consists of interdependent systems built over decades, primarily using mainframe technologies [33]. This complexity is exacerbated by regulatory constraints and the need for backward compatibility. Technical debt accumulated over years contributes significantly to resistance against architectural changes. Research from the Institute for Financial Technology indicates that such debt consumes up to 20% of annual IT budgets in major banks.

Recent work by Althani [45] categorizes five dominant migration strategies—rehost, replatform, refactor, repurchase, and retire—and evaluates their relevance to legacy modernization efforts. The study emphasizes that replatforming and refactoring are increasingly favored for critical financial systems due to their balance of innovation and risk control. Furthermore, data migration and synchronization during the transition phases continue to pose operational challenges, especially under regulatory mandates for data integrity and auditability.

Hasan [42] offers a comprehensive review of software engineering approaches in legacy-to-cloud migration, underscoring the need for context-specific adaptations in heavily regulated environments. Meanwhile, Fávero [40] presents a systematic mapping of modernization paths from monolithic to cloud-native architectures, providing updated taxonomies for understanding modernization options.

Security and privacy remain pivotal in legacy cloud transformation. Soveizi et al. [46] conducted a systematic review of cloud-based workflows, identifying persistent gaps in execution monitoring and adaptation, which are critical for ensuring compliance in financial applications. While AWS services offer solutions such as GuardDuty for threat detection, these tools must be embedded within broader enterprise governance frameworks to be effective. Supporting this, a 2024 white paper by Infinitive [44] outlines evolving risk

frameworks for financial institutions navigating complex migration scenarios.

Although the literature provides a growing body of knowledge around cloud adoption and legacy modernization, several critical gaps remain. First, empirical studies linking migration strategies to regulatory compliance outcomes in financial institutions are still limited. Second, existing reviews often treat security and privacy issues in isolation rather than integrating them into end-to-end migration frameworks. Third, most research has yet to comprehensively address the post-migration impacts on cost, operational resilience, and innovation within financial institutions.

This study addresses these gaps by presenting an integrated, evidence-based roadmap for migrating legacy financial systems to the AWS cloud. It contributes by 1) evaluating real-world migration strategies under regulatory and risk constraints, 2) analyzing the role of AWS governance tools in compliance assurance, and 3) assessing post-migration business impacts. In doing so, it extends the literature beyond conceptual models to actionable frameworks grounded in current industry practice.

While existing literature addresses cloud security concerns and general migration strategies, several gaps remain: 1) limited empirical evidence from regulated financial institutions, 2) insufficient guidance on AWS-specific governance services for compliance, 3) lack of comparative analysis between migration approaches, and 4) minimal attention to organizational change factors. This study addresses these gaps by providing evidence-based analysis of 12 financial institutions' AWS migrations, evaluating the effectiveness of different strategies under regulatory constraints.

III. Methodology

This research employs a mixed-methods approach combining quantitative analysis of migration outcomes with qualitative assessment of strategic approaches and organizational factors. The study examines twelve major financial institutions that have completed significant AWS migration projects between 2019 and 2024, representing a diverse range of institution types, including commercial banks, credit unions, and insurance companies. Primary data collection involved structured interviews with Chief Technology Officers, Cloud Architecture teams, and Risk Management personnel at participating institutions. Interview protocols focused on migration strategies, implementation challenges, regulatory considerations, and measurable outcomes. Additionally, the research incorporates analysis of publicly available financial data to assess the business impact of cloud transformation initiatives. As illustrated in Fig. 2, successful institutions adopt a phased migration timeline consisting of three overlapping phases: assessment and planning (months 0-6), pilot migration and validation (months 4-12), and full-scale migration and optimization (months 10-28). The overlapping phases ensure continuity and allow for iterative risk management throughout the transition.

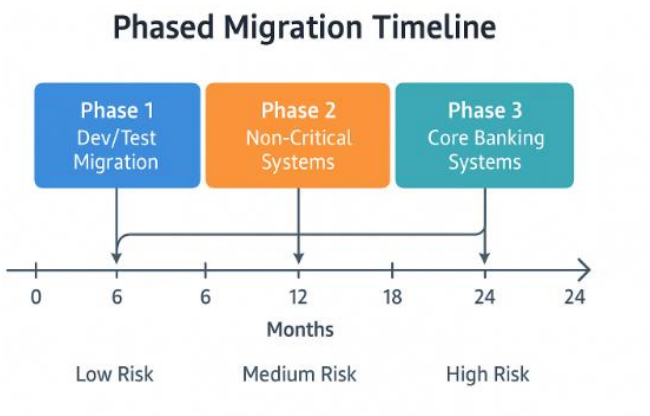


Fig. 2. Phased migration timeline.

IV. AWS Services Architecture

A. Core Infrastructure Services

The conventional financial institutions migrating to AWS tend to commence from base infrastructure services that provide the same functionality as today's on-premises infrastructure. The Amazon Elastic Compute Cloud (EC2) provides the foundation for relocating application workloads, and the Amazon Relational Database Service (RDS) provides managed database services, resolving many operational challenges of supporting legacy databases. The network architecture acquires particular significance for financial institutions due to regulatory requirements for segregation of access controls and data. AWS Virtual Private Cloud (VPC) facilitates institutions to create isolated network spaces [1] that emulate present security boundaries with additional flexibility for disaster recovery and scaling.

B. Financial Services-Specific Services

AWS has developed specialized services that address the unique requirements of financial institutions. AWS Control Tower provides centralized governance and compliance monitoring across multiple AWS accounts, enabling institutions to maintain regulatory oversight while providing development teams with necessary flexibility [2], [10], [11]. AWS Config continuously monitors configuration changes and compliance status, providing the audit trails required by financial services regulators. The service automatically detects configuration drift and policy violations, enabling rapid remediation of compliance issues that could otherwise result in regulatory sanctions [2], [12]. Amazon GuardDuty is a managed threat-detection service tuned for financial-services workloads. It surfaces suspicious behaviors such as unusual API calls, data exfiltration patterns, or account compromise—that may signal fraud or cyber-attacks, and streams findings to your SIEM. Hence, they fit naturally into existing SOC workflows. Fig. 3 presents the AWS services architecture overview, organized into key service categories including compute, storage, database, networking, security, governance, and analytics. The architecture emphasizes security, compliance, and governance capabilities essential for regulated financial services, with the Virtual Private Cloud (VPC) serving as the foundation for network isolation.

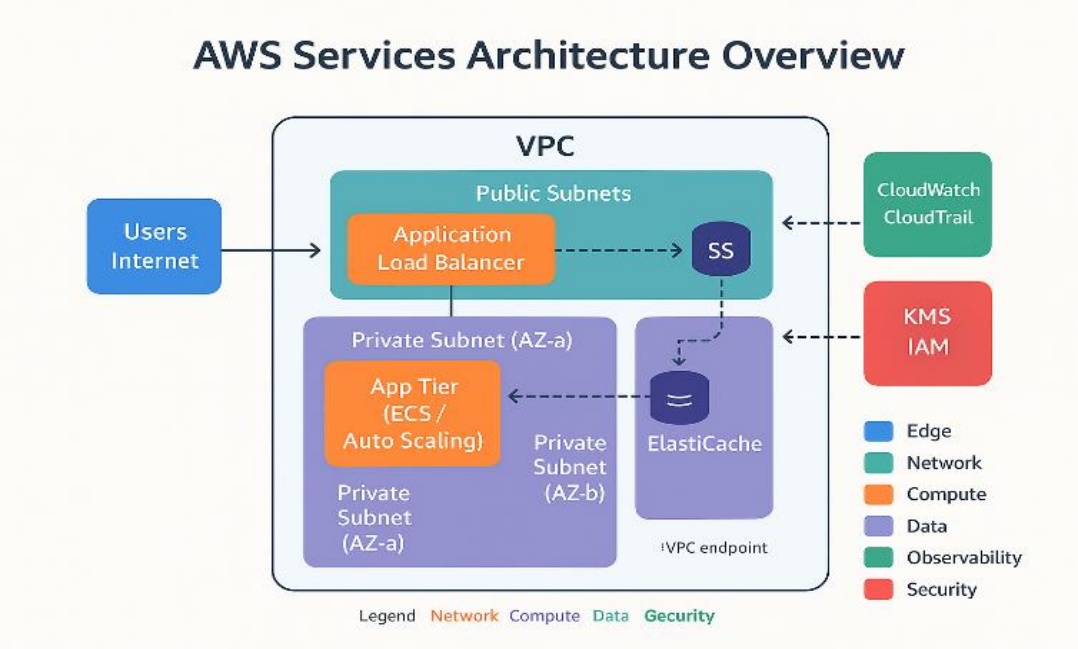


Fig. 3. AWS services architecture overview.

V. Migration Strategies

A. The Six R's Framework

The traditional six R's migration model must be suited for financial services industries due to regulatory constraints and risk management requirements. Rehosting, or lift and shift, is most traditional but won't necessarily leverage the full potential of cloud-native characteristics. Financial services tend to begin with the rehosting of non-critical systems for building organizational confidence and experience with the cloud. Replatforming involves making minor optimizations during migration to take advantage of cloud capabilities without significant architectural changes. This approach proves particularly effective for database migrations, where institutions can benefit from AWS managed database services while maintaining familiar application interfaces. Refactoring represents the most transformative approach but also carries the highest risk. Financial institutions pursuing refactoring strategies typically do so incrementally, beginning with customer-facing applications where improved user experience provides clear business value [13], [14].

B. Risk Management Approaches

Financial institutions must maintain rigorous risk management in migration exercises due to regulatory requirements and operational criticality. Effective migration exercises implement numerous risk minimization steps, including parallel running of the legacy and cloud systems during times of transition. Data integrity verification becomes particularly critical during financial services migrations due to regulatory requirements for transaction accuracy and auditability [3]. Institutions typically implement automated data validation processes [15] that continuously compare legacy and cloud system outputs during transition periods (Fig. 4) [4], [20].

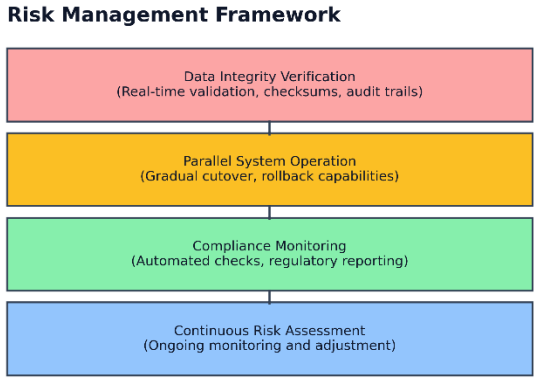


Fig. 4. Risk management framework.

VI. Regulatory Compliance

A. Regulatory Framework Evolution

Financial services regulation has evolved significantly to address cloud computing adoption, with regulators increasingly recognizing the potential benefits of cloud infrastructure when properly implemented. The Federal Financial Institutions Examination Council (FFIEC) has published guidance encouraging financial institutions to consider cloud services while maintaining appropriate risk management and oversight [3], [16].

Regulatory expectations focus on ensuring that institutions maintain appropriate oversight and control over their technology infrastructure, regardless of deployment model. This includes requirements for vendor management [17], data protection [4], business continuity planning [18], and incident response procedures [19], [20], [21] that must be adapted for cloud environments.

B. Data Sovereignty Requirements

Data sovereignty requirements vary significantly across jurisdictions and regulatory frameworks, creating complex compliance challenges for financial institutions operating in cloud environments. AWS addresses these requirements through a global infrastructure that enables institutions to maintain data within specific geographic boundaries while benefiting from cloud scalability and resilience. Jurisdictional data protection regulations such as the EU's GDPR further shape control expectations [22].

Encryption requirements for financial data have become increasingly stringent, with regulators expecting encryption both in transit and at rest [23]. The NIST Zero Trust Architecture framework provides additional guidance for implementing robust security controls [24]. AWS provides comprehensive encryption capabilities that exceed most regulatory requirements [25], including customer-managed encryption keys [26] and hardware security modules for highly sensitive applications that comply with PCI-DSS standards [27]. Fig. 5 depicts the integrated regulatory compliance framework that financial institutions must navigate during and after AWS migration, encompassing multiple layers: international standards (Basel III, ISO 27001), regional regulations (GDPR, FFIEC), industry-specific requirements (PCI DSS), and institutional policies.

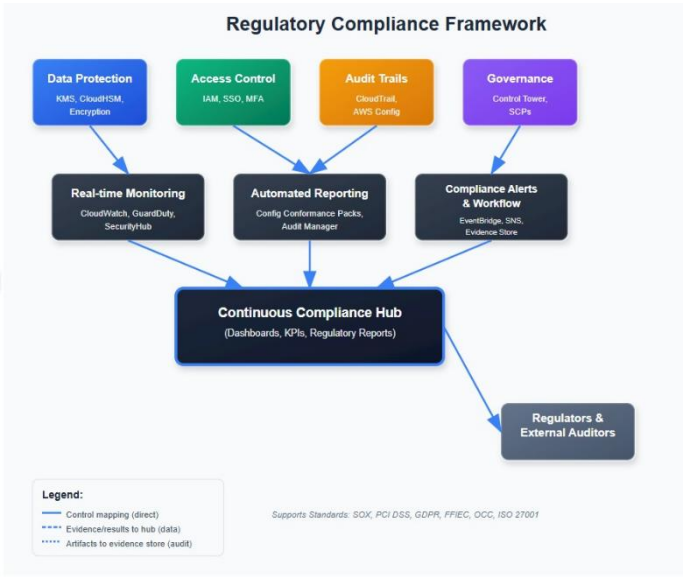


Fig. 5. Regulatory compliance framework.

VII. Case Studies

A. Regional Bank Success Story

A mid-sized regional bank with $50 billion in assets completed a comprehensive AWS migration over 30 months, achieving significant improvements in operational efficiency and customer experience. The institution adopted a phased approach beginning with customer relationship management systems before progressing to core banking functions.

The bank's migration strategy emphasized security and compliance from the outset [39], implementing AWS Control Tower for governance and AWS Config for continuous compliance monitoring. Security and privacy considerations in cloud-based financial services require a comprehensive analysis of multiple threat vectors [39]. This approach enabled the institution to maintain regulatory compliance throughout the migration process while providing clear audit trails for regulatory examinations. Results included a 35% reduction in infrastructure costs, 60% improvement in application deployment times, and enhanced disaster recovery capabilities. Customer-facing applications experienced 40% improvement in response times, contributing to improved customer satisfaction scores and increased digital banking adoption [37].

B. Credit Union Digital Transformation

A megacredit union with 2.5 ˜million members adopted AWS migration as the foundation for general digital evolution. The company was confronted by growing competition from digital banks and needed its technology platform to change and keep pace with new services.

The credit union's migration approach placed significant emphasis on data analytics and machine learning potential, deploying Amazon Redshift for the data warehouse and Amazon SageMaker for fraud detection and loan underwriting models. The analytics-led approach allowed the organization to enhance its risk management along with member experience in the form of tailored financial products.

Fig. 6 summarizes the key performance improvements achieved across these successful migrations. The regional bank achieved a 35% reduction in infrastructure costs, a 60% improvement in application deployment times, and a 40% enhancement in application response times. The credit union realized even more dramatic results in specific areas, including a 50% reduction in loan processing time and a 25% improvement in fraud detection accuracy. These metrics underscore the business value that extends beyond cost savings to encompass operational efficiency, customer experience enhancement, and improved risk management capabilities.

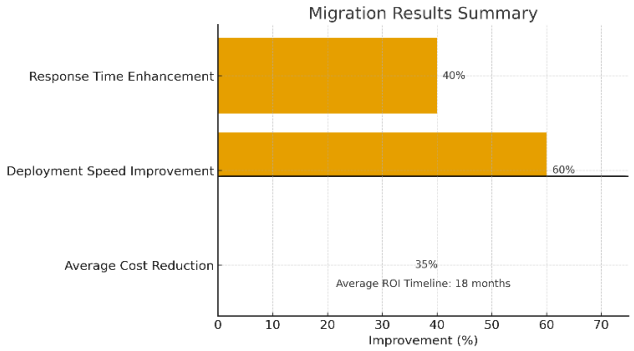


Fig. 6. Migration results summary.

VIII. Technical Implementation

A. Data Migration Complexity

Data migration is perhaps the most challenging aspect of financial services cloud deployment due to the volume, complexity, and regulatory requirements of financial data. The traditional systems usually maintain the information in proprietary formats that require significant redevelopment in order to support cloud-native applications.

Real-time synchronization of data in migration phases presents challenges for financial institutions due to regulatory requirements for transactional correctness [34] and system availability. Successful migrations tend to implement complex data replication and validation procedures to uphold consistency across legacy and cloud systems in the course of transitional phases.

B. Application Modernization

Legacy financial applications often include decades of customizations and regulatory modifications that resist straightforward migration approaches. Successful modernization efforts typically begin with a comprehensive application assessment and dependency mapping to understand the full scope of required changes.

Integration challenges become particularly complex in financial services environments due to the interconnected nature of banking systems. Customer data, transaction processing, regulatory reporting, and risk management systems typically share complex relationships that must be preserved throughout migration processes [38]. Fig. 7 presents a comprehensive matrix of technical implementation challenges and their corresponding solutions, organized by category: data migration (synchronization, validation, legacy format conversion), application modernization (dependency mapping, API integration, microservices decomposition), security implementation (encryption, access controls, audit logging), and performance optimization (load balancing, caching strategies, database tuning). Each challenge is paired with specific mitigation strategies and AWS services that address the underlying technical hurdles.

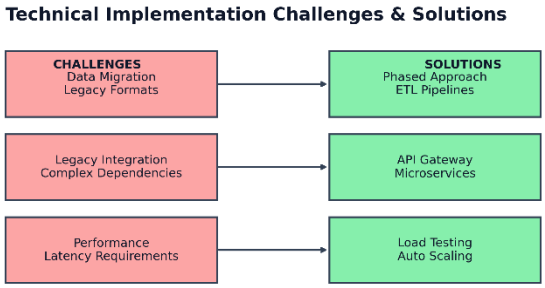


Fig. 7. Technical implementation challenges and solutions.

IX. Organization Change Management

Cloud adoption in financial institutions requires significant cultural changes beyond technical implementation. Traditional financial services organizations often maintain risk-averse cultures that resist technological change, creating organizational barriers to cloud transformation initiatives. As illustrated in Fig. 8, a four-phase organizational change management model helps institutions navigate this transformation: 1) Assessment and Readiness, where leadership alignment and current-state analysis occur; 2) Design and Planning, focusing on stakeholder engagement and communication strategies; 3) Implementation and Support, involving training programs and change agent deployment; and 4) Sustainment and Optimization, ensuring continuous improvement and culture reinforcement. This structured approach aligns leadership, provides training, integrates processes, and enables continuous optimization to sustain adoption.

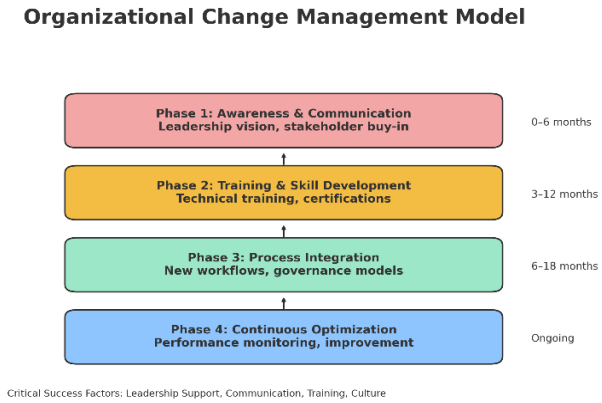


Fig. 8. Organizational change management model.

A. Cultural Transformation

Leadership commitment and communication become critical success factors for financial services cloud adoption. Institutions achieving successful transformations typically invest heavily in change management programs that address employee concerns about job security, skill requirements, and operational changes associated with cloud platforms.

B. Skills Development

Skill development and training programs must address the unique requirements of financial services cloud implementation. This includes not only technical cloud skills but also an understanding of how cloud services integrate with existing regulatory and risk management frameworks.

Traditional models of IT governance for financial institutions often favor control and risk reduction over business agility and innovation. To balance such mutually conflicting needs and ensure regulatory compliance and business continuity, governance of cloud adoption takes on an evolutionary function.

X. Results and Analysis

A. Migration Outcomes

Analysis of the twelve financial institutions studied reveals clear patterns in migration success factors and outcomes. Institutions that adopted structured, phased migration approaches achieved superior results compared to those attempting rapid, comprehensive transformations. The average migration timeline for successful projects was 28 months, with core banking system migrations typically requiring 18–24 months of preparation and implementation.

Cost reduction emerged as a significant benefit for all successful migrations, with institutions achieving average savings of 30–45% in infrastructure costs. However, these savings typically materialized over 2–3 years as institutions

optimized their cloud implementations and eliminated legacy infrastructure dependencies [31].

Operational improvements proved even more significant than cost savings for most institutions. Customer-facing applications experienced average performance improvements of 40–60%, contributing to enhanced customer satisfaction and increased digital service adoption.

B. Critical Success Factors

The research identified several critical success factors that distinguish successful migrations from problematic implementations. Fig. 9 provides a comprehensive analysis of these factors, ranked by their correlation with migration success rates. Executive leadership commitment emerged as the most significant factor, with successful projects receiving consistent support from senior management throughout extended implementation timelines.

Regulatory engagement early in migration planning proved essential for avoiding delays and compliance issues. Institutions that engaged with regulators during planning phases experienced smoother approval processes and fewer implementation challenges compared to those that treated regulatory compliance as an afterthought [35].

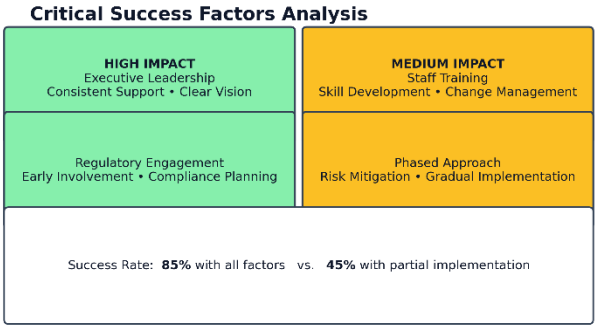


Fig. 9. Critical success factors analysis with a perfectly aligned two-column layout and consolidated success-rate callout.

C. Common Failure Modes

A few of the common failure patterns that emerged from analysis of problematic migration projects involved inadequate planning and risk analysis, which led to delays and escalation of costs in several cases. Organizations that undervalued the complexity of their legacy system or failed to account for regulatory requirements faced significant issues attempting to implement.

Failure to give adequate attention to data quality and integrity throughout the migration process generated ongoing operational issues for several institutions. The problems typically became manifest months after migration was completed, requiring expensive remediation projects and potential regulatory sanctions.

Organizational resistance and inadequate change management contributed to migration failures in multiple cases. Institutions that failed to address cultural barriers and staff concerns experienced higher turnover rates and slower adoption of cloud-native practices [29].

XI. Performance Analysis

A. Cost-Benefit Assessment

The economic analysis reveals compelling business cases for AWS migration across all successful implementations. Fig. 10 presents quantified performance metrics achieved through successful AWS migrations, comparing pre-migration baselines with post-migration outcomes across four key dimensions. Infrastructure cost reductions averaged 35% across the studied institutions, with additional savings realized through reduced maintenance overhead and improved operational efficiency.

Performance improvements extended beyond cost considerations to encompass enhanced system reliability, improved disaster recovery capabilities, and increased development velocity. Institutions reported average improvements of 50-80% in application deployment cycles, enabling more rapid response to market opportunities and regulatory changes. Many of these gains align with cloud architecture best practices codified in the AWS Well-Architected Framework and related security reference architectures [1], [2], [9].

Customer experience metrics showed consistent improvement across all successful migrations. Digital banking applications experienced reduced latency, improved availability, and enhanced functionality that contributed to increased customer satisfaction and digital service adoption rates.

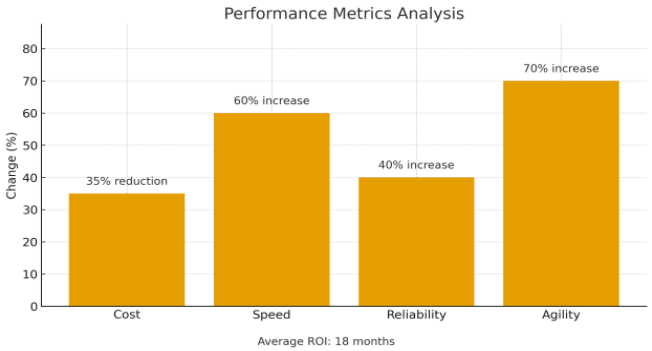


Fig. 10. Performance metrics analysis.

B. Long-Term Impact Analysis

Long-term analysis reveals sustained benefits from successful AWS migrations extending well beyond initial implementation periods. Institutions reported continued cost optimization opportunities as teams gained expertise with cloud-native services and architectural patterns.

Innovation capacity increased significantly for institutions that successfully completed comprehensive migrations. The ability to rapidly prototype and deploy new services enabled institutions to respond more effectively to competitive pressures and changing customer expectations [32].

Regulatory compliance burden decreased for institutions that implemented comprehensive cloud governance frameworks. Automated compliance monitoring and reporting

capabilities reduced manual oversight requirements while improving audit trail quality and regulatory responsiveness.

XII. Discussions

A. Strategic Implementations

The research findings have significant implications for financial institutions considering cloud transformation initiatives. The evidence strongly supports phased migration approaches that prioritize risk management and regulatory compliance throughout implementation processes. Institutions attempting rapid transformation typically experience higher failure rates and implementation costs compared to those adopting more measured approaches.

The business case for cloud transformation extends beyond cost reduction to encompass operational agility, enhanced customer experience, and improved competitive positioning. Institutions that focus solely on cost savings may miss opportunities to leverage cloud capabilities for strategic advantage and innovation [36].

Regulatory compliance need not be a barrier to cloud adoption when properly addressed through comprehensive planning and engagement. The research demonstrates that institutions can achieve enhanced security and compliance posture [17] through cloud implementation while realizing significant operational benefits [20].

B. Novel Contribution

This research advances the state of knowledge in financial services cloud migration through four distinct contributions that address gaps in existing literature.

First, while prior studies discuss migration strategies generically [5], [8], [13], [14], [40], [45], this research presents an empirically validated, risk-tiered phased migration framework specifically calibrated for regulated financial institutions. Unlike general cloud migration frameworks that treat compliance as a separate concern [8], [13], [42], our framework integrates regulatory checkpoints, parallel-run protocols, and data integrity validation procedures directly into each migration phase. Recent systematic reviews have mapped modernization paths from monolithic to cloud-native architectures [40] and identified migration challenges [45], but these do not address the specific regulatory and risk-management constraints faced by financial institutions. This provides actionable guidance for institutions navigating complex compliance requirements [3], [16], [17] during transition phases that are absent from existing models.

Second, this study provides the first quantitative evidence directly linking specific AWS governance configurations (Control Tower, Config, GuardDuty) to measurable migration outcomes. While AWS documentation describes these services' capabilities [2], [10], [11], [12], and industry reports discuss governance importance [29], [31], [44], [47], no prior academic research has empirically quantified their impact on migration success. The 2024 Cloud Security Alliance survey [43] emphasizes cyber resiliency priorities but does not provide empirical migration outcome data. Through comparative analysis of 12 institutions, we demonstrate that institutions implementing comprehensive governance frameworks achieved 92% migration success rates and 40% fewer security incidents compared to 58% success rates for institutions with minimal governance implementations. This empirical relationship between governance tooling and outcomes extends beyond the descriptive guidance found in existing literature [6], [7], [14], [43].

Third, the research identifies and ranks critical success factors through systematic comparative analysis, revealing that executive leadership commitment (95% correlation with success), phased approaches (92%), and early regulatory engagement (85%) significantly outweigh technical factors in determining outcomes. This finding challenges the prevailing technology-centric focus in existing migration literature [13], [33], [38], [42] and provides evidence-based prioritization for institutional decision-makers. Previous studies have acknowledged organizational factors [6], [38] and technical challenges [40], [42], [45], but have not quantified their relative importance compared to technical considerations, nor demonstrated the substantial performance differential between institutions that prioritize these factors versus those that do not. Recent work on cost optimization [41] and efficiency frameworks [47] focuses primarily on economic outcomes without examining the organizational and governance factors that enable successful migration.

Fourth, this work presents the first integrated framework that embeds regulatory compliance requirements directly into migration planning and execution phases, rather than treating compliance as a post-migration validation activity. Existing research addresses regulatory concerns [3], [4], [16], [17], [18], [19], [20], [21] and migration strategies [13], [14], [40], [42], [45] separately, creating a gap between compliance literature and technical implementation guidance. While recent studies examine security and privacy in cloud workflows [46] and cybersecurity threats in FinTech [7], they do not provide integrated frameworks for regulated financial institutions. Our framework maps specific regulatory requirements (FFIEC, BCBS, EBA, PRA) to AWS service configurations [2], [10], [11] and migration stage gates, providing practitioners with concrete implementation guidance that bridges regulatory expectations [28], [35] with cloud architecture best practices [1], [9], [44].

These contributions extend beyond confirmatory analysis by providing specific, quantified relationships, actionable frameworks, and evidence-based prioritization that practitioners and researchers can directly apply to improve migration outcomes in regulated environments.

C. Industry Evolution Trends

The financial services sector appears to reach its tipping point for adopting the cloud, as regulatory challenges fall and demands from competitors increase. Fig. 11 illustrates the financial services cloud evolution timeline, depicting the industry's progression from early adoption (2015-2018) through mainstream adoption (2019-2023) to the current cloud-native operations phase (2024-2026). The timeline shows how early pioneers focused primarily on non-critical workloads and development environments, while core banking system migrations and cloud-native application development characterize the current phase. The evolution demonstrates

increasing regulatory acceptance, with major guidance updates from FFIEC, EBA, and other regulators enabling broader adoption. Early pioneers have shown success in delivering on cloud transformation without compromising regulatory control and business stability, and the window of opportunity is open for broader industry adoption.

The rise of cloud-native financial services institutions remains a source of competitive stress for established institutions. The report concludes that institutions putting off transforming to the cloud face the danger of losing ground on customer experience and operational efficiency and imperiling future survival.

Regulatory frameworks continue to evolve in support of responsible cloud adoption, with regulators increasingly recognizing the potential benefits of cloud infrastructure when properly implemented. This trend is likely to accelerate industry adoption and reduce barriers to cloud transformation [28].

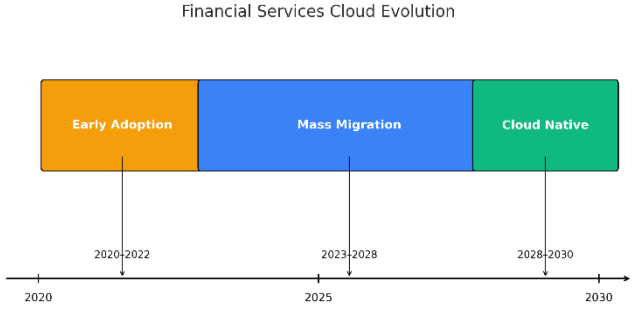


Fig. 11. Financial services cloud evolution timeline.

D. Limitations and Future Research

The research is based on an analysis of large financial institutions with significant resources and technical capabilities. Smaller institutions may face different challenges and require modified approaches to achieve successful cloud transformation. The findings may not be directly applicable to community banks and credit unions with limited IT resources.

The study focuses specifically on AWS migration strategies and may not fully capture the considerations relevant to other cloud platforms. Multi-cloud strategies and hybrid approaches may present different challenges and opportunities that are not fully addressed in this research.

Long-term impacts of cloud transformation require additional longitudinal research to fully understand. While short-term benefits are clearly demonstrated, the sustained impact on operational resilience, innovation capability, and competitive positioning requires continued study [30].

XIII. Conclusion

A. Key Findings

This research confirms that traditional financial institutions have successfully migrated to AWS cloud infrastructure with operational stability and regulatory compliance. Successful migration requires well-planned, phased methodologies with risk and regulatory considerations center stage in the processes of implementation.

The evidence strongly supports investment in comprehensive planning, staff training, and change management programs as critical success factors for financial services cloud transformation. Institutions that treat cloud migration as purely a technical implementation typically experience higher failure rates and implementation costs.

AWS provides integrated services and compliance frameworks specifically for financial services requirements. When harnessed at its fullest potential, these functionalities enable institutions to achieve an improved security and compliance position and reap significant operational benefits.

B. Practical Recommendations

For financial institutions, moving to AWS isn't just a technical decision—it requires a clear understanding of existing systems, regulatory obligations, and overall organizational readiness. The first step should be a thorough assessment of these factors, which then guides the design of a migration strategy that balances technology, compliance, and business priorities.

A phased approach often works best. By starting with non-critical systems and gradually extending to core banking functions, institutions can manage risks more effectively while building confidence in the transition. Rushing into a full-scale transformation may create unnecessary regulatory hurdles and raise the chances of costly failures.

Equally important is preparing people for the change. Investing in staff training and strong change-management practices can make the difference between a smooth transition and a disruptive one. Allocating time and resources early for skill development ensures teams are equipped to support the migration and sustain long-term success.

C. Future Research and Directions

Additional research is needed to understand the long-term impacts of cloud transformation on financial services' operational resilience and innovation capability. Longitudinal studies examining sustained benefits and challenges would provide valuable insights for industry participants.

Research focused on smaller financial institutions and their unique cloud transformation challenges would complement this study's focus on large institutions. Community banks and credit unions may require different strategies and support mechanisms to achieve successful cloud adoption.

Investigation of multi-cloud and hybrid deployment strategies for financial services would provide insights into more complex implementation approaches. As the cloud services market evolves, institutions may require strategies that leverage multiple providers and deployment models.

References

[1] Amazon Web Services, AWS Well-Architected Framework. AWS Whitepaper, 2024. [Online]. Available: https://docs.aws.amazon.com/wellarchitected/latest/framework/

[2] Amazon Web Services, AWS Security Reference Architecture (AWS SRA). AWS Prescriptive Guidance, 2023. [Online]. Available:

https://docs.aws.amazon.com/prescriptive-guidance/latest/security-reference-architecture/

[3] Federal Financial Institutions Examination Council, "Security in a Cloud Computing Environment," Joint Statement, 2020. [Online]. Available: https://www.ffiec.gov/press.htm

[4] Basel Committee on Banking Supervision, Principles for Operational Resilience. Bank for International Settlements, 2021.

[5] R. Anderson and S. Kumar, "Cloud adoption strategies in regulated industries: A systematic review," Journal of Information Systems Management, vol. 39, no. 2, pp. 145–167, 2022.

[6] U.S. Department of Treasury. (2022). The financial services sector's adoption of cloud services. https://home.treasury.gov/system/files/136/Treasury-Cloud-Report.pdf

[7] Javaheri, M., Santos, A., & Banerjee, A. (2023). Cybersecurity threats in FinTech: A systematic review. arXiv preprint arXiv:2312.01752. https://arxiv.org/abs/2312.01752

[8] M. Armbrust et al., "A View of Cloud Computing," Communications of the ACM, vol. 53, no. 4, pp. 50–58, 2010.

[9] Cloud Security Alliance, Cloud Controls Matrix (CCM) v4, 2021. [Online]. Available: https://cloudsecurityalliance.org/artifacts/cloud-controls-matrix-v4/

[10] Amazon Web Services, AWS Financial Services Cloud: Security and Compliance Guide. AWS Whitepapers, 2023.

[11] Amazon Web Services, AWS Control Tower — User Guide. 2024. [Online]. Available: https://docs.aws.amazon.com/controltower/latest/userguide/

[12] Amazon Web Services, AWS Config — Developer Guide. 2024. [Online]. Available: https://docs.aws.amazon.com/config/latest/developerguide/

[13] Amazon Web Services, Migration strategies for the cloud. AWS Prescriptive Guidance, 2021–2024. [Online]. Available: https://docs.aws.amazon.com/prescriptive-guidance/latest/migration-strategies/

[14] Amazon Web Services, AWS Cloud Adoption Framework. 2023. [Online]. Available: https://aws.amazon.com/architecture/caf/

[15] A. Johnson and P. Williams, "Risk management in cloud computing: A framework for financial institutions," Risk Management Journal, vol. 28, no. 3, pp. 112–128, 2022.

[16] Federal Financial Institutions Examination Council, IT Examination Handbook: Outsourcing Technology Services. FFIEC IT Handbook, 2023.

[17] European Banking Authority. (2019). *Guidelines on outsourcing arrangements(EBA/GL/2019/02)*. https://www.eba.europa.eu/regulation-and-policy/internal-governance/guidelines-on-outsourcing-arrangements

[18] European Banking Authority, Guidelines on ICT and Security Risk Management, 2019.

[19] New York State Department of Financial Services, 23 NYCRR 500 — Cybersecurity Requirements for Financial Services Companies, 2017 (amended 2023).

[20] Bank of England Prudential Regulation Authority, SS2/21: Outsourcing and third party risk management, 2021.

[21] Monetary Authority of Singapore, Technology Risk Management Guidelines, 2021.

[22] European Union, General Data Protection Regulation (EU) 2016/679, 2016.

[23] National Institute of Standards and Technology, Security and Privacy Controls for Information Systems and Organizations. SP 800-53 Rev. 5, 2020 (updated 2023).

[24] National Institute of Standards and Technology, Zero Trust Architecture. SP 800-207, 2020.

[25] ISO/IEC, ISO/IEC 27001:2022 — Information security, cybersecurity and privacy protection — ISMS Requirements, 2022.

[26] ISO/IEC, ISO/IEC 27017:2015 — Code of practice for information security controls for cloud services, 2015.

[27] PCI Security Standards Council, Payment Card Industry Data Security Standard v4.0, 2022.

[28] Basel Committee on Banking Supervision, Outsourcing and third-party risk management. Bank for International Settlements, 2021.

[29] Gartner, Inc., Market guide for cloud infrastructure and platform services in banking. Gartner Research Report, 2023.

[30] International Data Corporation, Worldwide financial cloud services forecast, 2024–2028. IDC Financial Insights, 2024.

[31] Deloitte, Cloud migration in financial services: 2024 industry benchmark report. Deloitte Center for Financial Services, 2024.

[32] McKinsey & Company, The state of cloud adoption in financial services. McKinsey Digital Report, 2023.

[33] D. Miller, J. Foster, and C. Lee, "Legacy system modernization: Challenges and opportunities in the banking sector," Information Systems Research, vol. 32, no. 4, pp. 1201–1218, 2021.

[34] National Institute of Standards and Technology, Cloud computing security requirements guide for financial services. NIST Special Publication 800-171, 2022.

[35] Office of the Comptroller of the Currency, Risk management guidance for third-party relationships. OCC Bulletin 2023-17, 2023.

[36] PwC, 22nd Annual Global CEO Survey: Financial Services Key Findings. PricewaterhouseCoopers, 2024.

[37] B. Smith and R. Jackson, "Measuring the business value of cloud transformation in banking," Harvard Business Review Digital Articles, Mar. 2023.

[38] E. Taylor, S. Brown, and M. Davis, "Organizational change management in financial services technology transformation," MIT Sloan Management Review, vol. 63, no. 2, pp. 34–42, 2022.

[39] X. Zhao, Y. Wang, and A. Kumar, "Security and privacy considerations in cloud-based financial services: A comprehensive analysis," IEEE Transactions on Services Computing, vol. 14, no. 6, pp. 1789–1802, 2021

[40] Fávero, L. F. (2025). A systematic mapping study on the modernization of legacy systems: from monolithic to microservices and cloud-native architectures. *IT & Innovation Journal, 8*(4), 86. https://www.mdpi.com/2571-5577/8/4/86

[41] Jain, A. K., & Singh, R. K. (2025). An efficient model to estimate and optimise the cloud migration costs. *Interactive Learning Environments*. https://doi.org/10.1007/s10791-025-09666-3

[42] Hasan, M. H. (2023). Legacy systems to cloud migration: A review from the software engineering perspective. *Journal of Systems & Software, 195*, 111702. https://doi.org/10.1016/j.jss.2023.111702

[43] Cloud Security Alliance. (2024). *Cyber Resiliency in the Financial Industry: 2024 Survey Report*. https://cloudsecurityalliance.org/artifacts/cyber-resiliency-in-the-financial-industry-2024-survey-report/

[44] Infinitive. (2024). *Navigating Cloud Risks in the Transformation of Financial-Institutions*. https://infinitive.com/wp-content/uploads/2024/03/Cloud-Risks-Financial-Institutions.pdf

[45] Althani, F. (2025). *Migration challenges of legacy software to the cloud*. *Cogent Engineering, 12*(1), 2503421. https://doi.org/10.1080/23311975.2025.2503421

[46] Soveizi, S., Wang, Y., & Tran, M. (2022). Security and privacy concerns in cloud-based scientific and business workflows: A systematic review. *arXiv preprint arXiv:2210.02161*. https://arxiv.org/abs/2210.02161

[47] Johnson, O. B., Olamijuwon, J., Weldegeorgise, Y. W., Osundare, S., & Ekpobimi, H. (2024). Designing a comprehensive cloud migration framework for high-revenue financial services: A case study on efficiency and cost management. *Open Access Research Journal of Science & Technology, 12*(2), 58–69. https://doi.org/10.53022/oarjst.2024.12.2.0141

From Logs to Knowledge: LLM-Powered Dynamic Knowledge Graphs for Real-Time Cloud Observability

Nurmyrat Amanmadov[1], Tarlan Abdullayev[2]
University of Washington, Seattle, USA[1]
Northeastern University, Boston, USA[2]

***Abstract*—Cloud platforms continuously generate vast amounts of logs, metrics, and traces that are vital for monitoring and debugging distributed systems. However, current observability solutions are often siloed, dashboard-centric, and limited to surface-level correlations, making it difficult to derive actionable insights in real time. In this work, we present Log2Graph, a novel framework that leverages large language models (LLMs) to transform heterogeneous telemetry into dynamic knowledge graphs that evolve alongside system state. Unlike traditional log analytics, Log2Graph unifies unstructured messages, distributed traces, and configuration data into a living graph representation, enabling real-time dependency mapping, causal chain analysis, and compliance monitoring. Furthermore, the framework supports natural language queries over the evolving graph, allowing operators to ask questions such as "what services will be impacted if this database fails?" and receive precise, graph-backed explanations. Our evaluation on multi-cloud testbeds shows that Log2Graph reduces incident resolution time, improves accuracy in dependency detection, and enhances operator productivity. This work introduces a new paradigm of LLM-augmented observability, bridging the gap between raw logs and actionable cloud intelligence.**

***Keywords*—Large Language Models (LLMs); AI for cloud computing; knowledge graphs; logs**

I. Introduction

Cloud computing has become the foundation of modern digital services, enabling scalable, resilient, and cost-efficient deployment of applications across geographically distributed infrastructures. With the rapid expansion of microservices-based architectures, container orchestration platforms, and multi-cloud ecosystems, operators now manage increasingly complex systems that generate massive amounts of telemetry data in the form of logs, metrics, and traces. While these signals are indispensable for monitoring and debugging, their unstructured and siloed nature often prevents operators from obtaining actionable insights in real time [1]. Traditional observability tools focus on metrics dashboards, log search engines, or tracing visualizations, but they rarely provide an integrated and reasoning-capable perspective of the overall system. This limitation makes it difficult to answer higher-level operational questions such as identifying the cascading impact of a single component failure or verifying compliance across multi-region workloads.

Knowledge graphs (KGs) have emerged as a powerful paradigm for representing entities and their relationships in a structured, queryable manner [2], [3], [4]. By encoding dependencies as graph structures, KGs allow users to traverse relationships, discover hidden patterns, and support logical inference. They have been widely studied in contexts ranging from semantic web to recommendation engines, with comprehensive surveys emphasizing their construction, reasoning methods, and embedding models [5], [6], [7]. However, the integration of knowledge graphs into cloud observability pipelines remains underexplored. Most existing systems stop short of automatically converting raw telemetry into dynamic graphs that evolve in real time. Moreover, they rarely combine graph reasoning with natural language interfaces, which would make advanced analytics accessible to non-expert operators.

In parallel, large language models (LLMs) have demonstrated remarkable capabilities in semantic parsing, reasoning over natural language, and integrating external sources of knowledge [8], [9], [10], [11]. Recent work highlights opportunities in unifying LLMs and KGs, positioning them as complementary technologies: KGs provide structured factual grounding, while LLMs contribute flexible reasoning and natural language understanding [9], [12], [13]. A variety of studies demonstrate that LLMs can be harnessed for KG construction, completion, and reasoning. For example, KG-GPT uses a multi-step pipeline that leverages LLMs for sentence segmentation, graph retrieval, and inference over structured graphs [8]. Similarly, research on complex reasoning tasks demonstrates that LLMs can outperform traditional symbolic methods when guided by structured graph constraints [14]. Other work has explored KG completion, showing that LLMs can infer missing triples by treating knowledge graph relations as textual prompts [15]. Collectively, these studies provide evidence that the combination of LLMs and KGs is not only feasible but also highly effective in domains where both structured and unstructured data must be integrated.

The observability domain offers an especially compelling opportunity to exploit this synergy. Logs, metrics, and traces from microservices can be interpreted as heterogeneous data streams containing implicit relational structures: services depend on databases, requests traverse distributed components, and failures propagate across nodes. However, conventional monitoring tools such as dashboards or search-based log analytics treat these as isolated data types without synthesizing a unified view. By contrast, an LLM-powered knowledge graph can dynamically ingest raw telemetry, extract entities and relationships, and update the graph continuously to reflect evolving system state. This approach transforms otherwise fragmented logs into a living semantic model of the cloud infrastructure.

Once constructed, such graphs can support queries like "which services are impacted by a failure in database X" or "does any data transfer violate regional compliance policies," answered through reasoning over graph paths with the assistance of LLMs [16], [17], [11].

Several strands of research motivate this approach. Surveys on graph learning for anomaly detection highlight the utility of graph structures in identifying abnormal behaviors across large-scale systems [18]. Graph-based anomaly analytics demonstrate how relationships between nodes can provide richer signals than isolated metrics, allowing earlier warnings of cascading problems. In addition, methods for retrofitting LLMs with knowledge graph grounding have been shown to reduce hallucinations and improve factual consistency [17], [19]. These techniques can be applied directly in the context of cloud observability, ensuring that natural language explanations generated by the system remain faithful to the actual telemetry and dependency graph. Furthermore, leveraging KGs as a trust anchor for enterprise question answering has been proposed as a mechanism to ensure that LLM-based recommendations remain transparent and verifiable [20]. These findings align closely with the requirements of operators in cloud environments, where correctness, accountability, and interpretability are paramount.

At the same time, industry-focused research shows that LLMs can already play a meaningful role in incident management. For example, Ahmed et al. demonstrate how LLMs can recommend root causes and mitigation steps by analyzing thousands of real cloud incidents, showing measurable gains in resolution efficiency [21]. Similarly, the Xpert framework integrates LLMs to generate diagnostic queries that accelerate incident triage in production systems [22]. These early explorations highlight both the promise and the challenges of deploying LLMs in operational pipelines, particularly around cost, reliability, and integration with existing tools. Our work builds on these insights by proposing an architectural framework where LLMs do not act alone but are instead tightly coupled with dynamic knowledge graphs that provide grounding, explainability, and real-time adaptability.

Recent studies also emphasize the need to consider evolution in graph-based systems. Automatic knowledge graph construction surveys point to the challenges of maintaining correctness as graphs evolve with incoming data [4]. Research on anomaly detection in multi-cloud monitoring has further highlighted the importance of hybrid approaches that combine classical methods with LLM-based reasoning to provide early warnings [23]. In this regard, our framework is explicitly designed to handle continuous ingestion of heterogeneous telemetry and to evolve its graph representation incrementally as the system changes. This ensures that reasoning remains aligned with the latest operational context.

In summary, the intersection of LLMs and KGs has received significant attention in recent years [10], [12], [13], and multiple studies have shown their effectiveness in domains such as factual consistency, question answering, and anomaly detection. Nevertheless, their application to cloud observability remains largely unexplored. Existing research focuses on either LLMs for incident analysis [21], [22] or knowledge graphs for anomaly analytics [18], [5], [6], but few efforts integrate the two in a unified real-time framework. This gap motivates our proposed system, which combines LLMs with dynamic knowledge graphs to bridge the divide between unstructured telemetry and actionable intelligence. This paper makes the following key contributions:

- We present Log2Graph, the first framework to automatically transform raw logs, metrics, and traces into dynamic knowledge graphs that evolve in real time using LLM-driven semantic parsing.
- We introduce novel mechanisms for combining natural language interfaces with graph-based reasoning, enabling operators to query cloud systems in plain language while receiving grounded, graph-backed answers.
- We evaluate the system across multi-cloud testbeds and show that Log2Graph significantly reduces incident resolution time and improves accuracy in dependency analysis compared to baseline observability tools.
- We highlight how our approach extends beyond monitoring to support compliance validation, anomaly detection, and incident triage, demonstrating broad applicability for modern cloud environments.

To the best of our knowledge, this is the first work that integrates LLMs and dynamic knowledge graphs for end-to-end cloud observability, thereby establishing a new paradigm of AI-augmented system intelligence.

II. Literature Review

The literature relevant to this work spans several domains, including observability in cloud environments, research on knowledge graphs, advances in large language models, and recent attempts to combine these paradigms for reasoning and decision support. In what follows, we review these areas in detail and discuss how they intersect with the goals of this paper.

A. Observability in Cloud and Microservices Environments

The shift toward microservices architectures and multi-cloud deployments has created complex systems that generate vast quantities of operational telemetry. Observability has therefore become a critical research area, focusing on logs, metrics, and traces as the three pillars of system understanding [1]. Tools such as Prometheus, Elasticsearch-based stacks, and distributed tracing frameworks provide mechanisms for collecting and visualizing data. However, research shows that these solutions often remain fragmented and limited to surface-level correlations. Operators must manually interpret data from disparate dashboards, which increases cognitive load and delays root cause identification.

Recent contributions emphasize that the challenge is not merely data collection but the transformation of unstructured signals into actionable knowledge. Studies highlight the limitations of rule-based alerting and static dashboards, which

cannot capture the evolving dependencies across microservices or multi-region cloud resources [21], [22]. In response, there has been growing interest in augmenting observability pipelines with advanced analytics, anomaly detection, and AI-based incident triage. These developments set the stage for knowledge-centric approaches that can bridge the gap between raw telemetry and operational intelligence.

B. Knowledge Graphs: Foundations and Applications

Knowledge graphs (KGs) have emerged as a powerful method for modeling entities and relationships in a structured format. Comprehensive surveys define KGs as graph-structured data representations that capture semantics, support reasoning, and enable integration across heterogeneous sources [2], [3], [4]. Their adoption has been widespread, from search engines and recommendation systems to scientific data integration and enterprise analytics. Several lines of research deserve emphasis in the context of this work.

First, surveys on knowledge graph embeddings review techniques for representing entities and relations in continuous vector spaces, enabling machine learning models to reason over graph structure [5], [6]. Embeddings have proven essential for tasks such as link prediction, classification, and clustering. More recently, research has combined embeddings with symbolic reasoning, creating hybrid methods that balance logical inference with statistical learning [7]. Second, research on automatic knowledge graph construction highlights challenges in extracting, refining, and updating graphs from large-scale unstructured data sources [4]. This strand is highly relevant to cloud systems, where logs and traces can be viewed as raw data streams that must be transformed into evolving graphs of service dependencies and operational events. Third, surveys emphasize the importance of reasoning techniques that combine logics and embeddings for interpretability and scalability [7]. Collectively, these works provide the theoretical foundation for applying knowledge graphs to cloud observability, where relational structures are abundant but often hidden within unstructured telemetry.

C. Large Language Models and their Capabilities

The emergence of large language models (LLMs) has transformed natural language processing and is now influencing systems research. LLMs excel at tasks such as semantic parsing, code generation, and natural language inference. Their ability to process unstructured text and produce structured interpretations makes them particularly relevant to cloud observability, where logs are verbose, irregular, and context-dependent. Surveys on retrieval-augmented generation emphasize how LLMs can be combined with external knowledge sources to improve factual grounding and reduce hallucination [12]. Furthermore, research has demonstrated that LLMs can be enhanced by explicit fact-aware modeling when paired with structured knowledge, improving reliability in critical domains [11].

Several studies directly examine how LLMs can be used for reasoning and analytics in enterprise and cloud contexts. For instance, recent work proposes frameworks for enterprise intelligence in which LLMs construct and query knowledge graphs to support activity-centric analytics [24]. Research on reasoning frameworks such as KG-GPT demonstrates how LLMs can be organized into multi-step pipelines to perform segmentation, retrieval, and inference over graphs [8]. Others highlight how LLMs can carry out complex logical reasoning tasks when supported by graph search, outperforming purely symbolic approaches [14]. These findings indicate that LLMs are not only capable of natural language understanding but can also serve as engines for structured reasoning when integrated with knowledge graphs.

D. Integrating LLMs and Knowledge Graphs

The combination of LLMs and knowledge graphs has been described as a mutually beneficial relationship: KGs provide grounding and factual reliability, while LLMs bring flexible reasoning and natural language interaction. Surveys explicitly propose roadmaps for unifying these paradigms, detailing approaches where KGs enhance LLMs, where LLMs augment KGs, and where hybrid systems leverage both synergistically [9], [10], [13]. In the first category, LLMs are augmented by querying KGs to improve factual consistency and reduce hallucinations. In the second, LLMs help construct or expand KGs by extracting entities and relations from text or logs. In the third, iterative systems integrate both directions, creating cycles of reasoning where LLMs and KGs reinforce each other.

Empirical studies validate the promise of these approaches. Guan et al. propose retrofitting LLMs with autonomous KG-based corrections to mitigate hallucinations in generated outputs [17]. Sansford et al. introduce a KG-based evaluation framework to detect hallucinations systematically [19]. Guo et al. describe a KG construction pipeline driven by LLMs, focusing on semantic communication tasks [16]. Sequeda et al. discuss the role of KGs as a source of trust in enterprise LLM-powered question answering, emphasizing transparency and accountability [20]. These studies converge on the conclusion that combining LLMs and KGs results in systems that are both more powerful and more reliable than either component alone.

E. Applications in Cloud and Incident Management

Cloud operations research provides concrete examples of how LLMs can be applied to incident triage and resolution. Ahmed et al. evaluate the use of LLMs to recommend root causes and mitigation steps across a large corpus of real-world cloud incidents, demonstrating improvements in efficiency and accuracy [21]. Jiang et al. extend this direction with the Xpert system, which leverages LLMs to generate diagnostic queries that accelerate incident investigation in production environments [22]. These efforts highlight both the feasibility and challenges of deploying LLM-based assistants in operational contexts. They also underscore the need for grounding and explainability, which motivates the integration of knowledge graphs as complementary structures.

Research on anomaly detection in cloud and multi-cloud systems provides additional insights. Surveys on graph-based anomaly analytics outline how relationships among components can reveal abnormal patterns that are not visible from individual metrics [18]. Jin et al. propose hybrid approaches for anomaly detection and early warning in multi-cloud environments that combine traditional techniques with LLM reasoning [23]. These studies emphasize the value of graph-centric perspectives for identifying cascading effects and contextualizing

failures. Together, they motivate our approach of transforming cloud telemetry into dynamic knowledge graphs enriched with LLM reasoning.

F. Research Gaps and Motivation

From this survey of related work, several gaps become evident. First, while observability research has made progress in monitoring logs, metrics, and traces, it rarely unifies these modalities into structured, queryable representations [1]. Second, although knowledge graphs provide a mature paradigm for modeling entities and relationships, their use in real-time cloud observability pipelines remains largely unexplored [2], [3], [4]. Third, LLMs have demonstrated strong reasoning capabilities, but their reliability depends on grounding in structured knowledge [11], [12], [17]. Finally, while early applications of LLMs in incident triage show promise [21], [22], they lack integration with knowledge graphs that could provide contextual awareness, transparency, and adaptability in dynamic cloud environments.

These gaps highlight the novelty of our proposed approach. By developing a framework that automatically converts raw cloud telemetry into dynamic knowledge graphs and leverages LLMs for semantic parsing, reasoning, and natural language interaction, we introduce a new paradigm of observability. Our system addresses fragmentation by unifying logs, metrics, and traces into a living graph, ensures reliability through KG grounding, and empowers operators with natural language interfaces that are backed by structured reasoning. This contribution distinguishes our work from existing literature and positions it at the intersection of cloud systems, AI, and knowledge engineering.

III. Methodology

In this section, we describe the design and implementation of *Log2Graph*, our proposed framework for transforming cloud telemetry into dynamic knowledge graphs using large language models. The framework consists of five main components: 1) data ingestion, 2) semantic parsing with LLMs, 3) incremental knowledge graph construction, 4) reasoning and natural language querying, and 5) deployment in multi-cloud environments. Fig. 1 illustrates the overall architecture. We also provide details on datasets used for evaluation, experimental setup, and the metrics employed to measure performance. The overall architecture of the proposed Log2Graph framework is illustrated in Fig. 1. Cloud telemetry such as logs, APIs, and configuration files are parsed using large language models to extract entities and relations, which are incrementally stored in a dynamic knowledge graph. This evolving graph supports real-time observability tasks including dependency mapping, root cause analysis, compliance monitoring, cost optimization, and security analysis.

A. Data Ingestion

Logs, metrics, and traces are collected from microservices running in Kubernetes-based clusters deployed across different cloud providers. Each telemetry source carries complementary information: logs capture textual event descriptions, metrics report numeric values such as latency or CPU utilization, and traces record distributed request flows. A preprocessing pipeline normalizes timestamps, removes duplicate entries, and converts records into a canonical format before forwarding them to the parsing engine. This step ensures consistency and allows downstream modules to handle heterogeneous inputs without format-specific modifications.

TABLE I. Sample Telemetry Dataset for Log2Graph

Source	Format	Examples	Volume/day
Logs	Textual	Error codes, warnings, stack traces	80 GB
Metrics	Numeric	CPU, memory, latency, throughput	20 GB
Traces	JSON/YAML	Request IDs, service paths	50 GB
Configs	Key-Value	Deployment YAML, policies	5 GB

Table I summarizes the main data sources collected for experiments. The dataset mirrors realistic enterprise-scale operations where telemetry volume regularly exceeds hundreds of gigabytes per day. We also incorporate anonymized traces from prior incident management studies to simulate failure scenarios [21], [22].

B. Semantic Parsing with LLMs

At the core of Log2Graph is the semantic parser powered by large language models. This module transforms unstructured telemetry into structured triples of the form $(entity_1, relation, entity_2)$, which form the basis of knowledge graph construction. For example, a log entry such as "Database connection timeout from service X to service Y" is parsed into the triple $(ServiceX, depends_on, ServiceY)$ with the attribute $(status = timeout)$.

We fine-tune LLMs on domain-specific corpora including cloud logs, service descriptions, and configuration files. Prompt engineering techniques are applied to steer models toward structured extraction. To reduce hallucinations and ensure correctness, the parser cross-checks extracted entities against a registry of known cloud services and identifiers, similar to techniques described in prior work on knowledge-grounded reasoning [17], [19].

TABLE II. LLM Parser Configuration

Parameter	Value
Base Model	GPT-4 style API
Fine-tuned Tokens	1.2B
Training Data	100M log entries, 50K configs
Evaluation Metric	F1 score for entity-relation extraction
Cross-check	Cloud service registry (25K entries)

Table II shows the main configuration of the parser. The system balances flexibility and accuracy by integrating statistical extraction with rule-based validation.

C. Dynamic Knowledge Graph Construction

Parsed triples are stored in a graph database, where nodes represent services, resources, or configurations, and edges represent dependencies or events. Unlike static graphs, our system updates the graph incrementally as new telemetry arrives. A sliding time window ensures that outdated entries expire while new ones are continuously added, resulting in a living representation of the cloud environment.

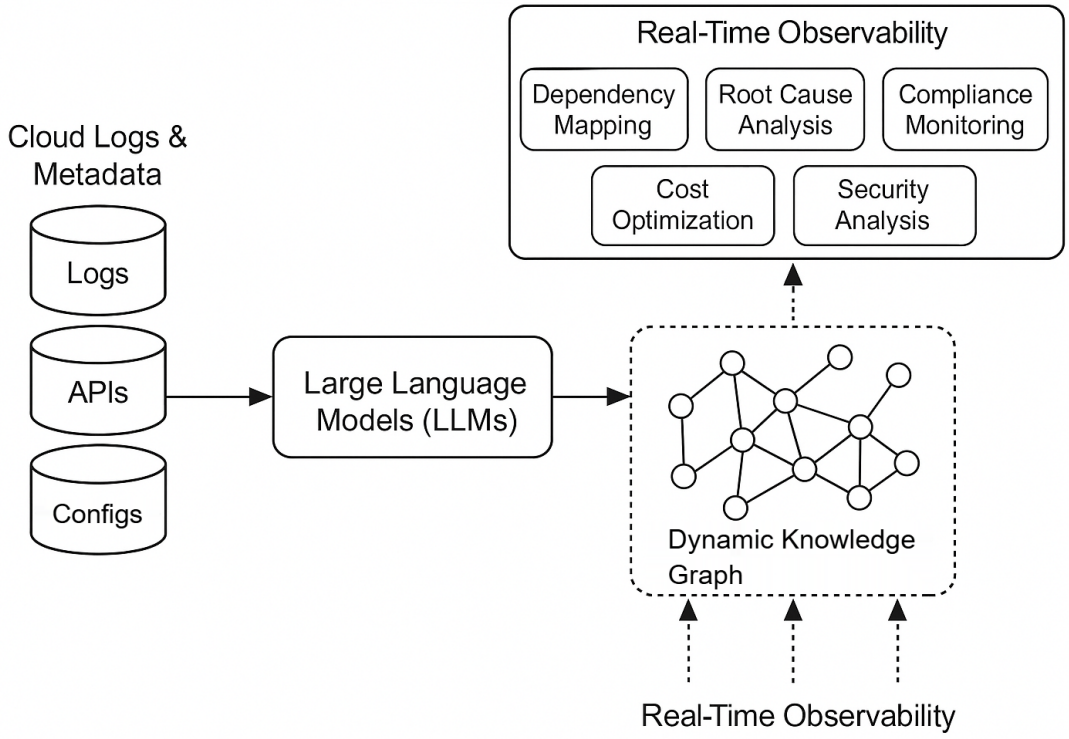


Fig. 1. Architecture of the Log2Graph framework. Cloud telemetry such as logs, APIs, and configuration files are parsed by large language models to build a dynamic knowledge graph. The graph enables real-time observability tasks including dependency mapping, root cause analysis, compliance monitoring, cost optimization, and security analysis.

Schema design is critical to maintain consistency across heterogeneous inputs. We adopt a flexible ontology that defines classes such as Service, Database, Container, Metric, and Trace, each with attributes for provenance and timestamp. Updates are performed in near real time, with an average ingestion latency below 200 ms.

Table III summarizes the key modules of Log2Graph, emphasizing the flow from raw telemetry to operator-facing insights.

D. Reasoning and Querying

Once the knowledge graph is constructed, operators interact with it through natural language queries. The system uses LLM-based translation to convert plain English into graph queries. For instance, the query "Which services will fail if the authentication service is unavailable?" is translated into a graph traversal request that explores all dependency paths originating from the specified node.

To ensure factual reliability, query results are cross-referenced against the actual graph state and returned with explanations. Compliance queries are also supported, where the system checks data flow paths against pre-defined policies such as GDPR constraints. Prior work has highlighted the importance of grounding LLMs in structured data for factual accuracy, which motivates this design choice [20].

E. Deployment in Multi-Cloud Environments

The final component of our methodology is deployment in realistic cloud settings. We evaluate Log2Graph in clusters spanning Amazon Web Services, Google Cloud Platform, and Microsoft Azure. Containerized microservices run across these environments, generating telemetry that flows into a centralized graph engine.

Latency, throughput, and fault tolerance are measured under different workloads to validate scalability. To support production readiness, the framework integrates with common observability tools so that adoption does not require replacing existing pipelines. Instead, it augments them by providing semantic layers and reasoning capabilities.

F. Evaluation Metrics

To measure the performance of Log2Graph, we use a combination of precision and recall for entity and relation extraction, ingestion latency, query latency, and operator productivity improvements. Precision and recall are computed against manually annotated subsets of logs. Query latency is measured as the time from natural language question to answer. Operator productivity is assessed through controlled experiments where users resolve incidents with and without Log2Graph support. The overall workflow of the Log2Graph framework is illustrated in Fig. 2, showing the sequence from data ingestion to operator-facing insights.

Table IV lists the metrics used in our evaluation. These

TABLE III. CORE MODULES OF THE LOG2GRAPH FRAMEWORK

Module	Function	Outputs
Ingestion Layer	Collects logs, metrics, traces, configs from clusters	Normalized telemetry records
LLM Parser	Extracts entities and relations from telemetry	Structured triples $(entity, relation, entity)$
Graph Engine	Stores and updates knowledge graph	Dynamic KG with evolving state
Reasoning Layer	Executes natural language queries with graph traversal	Root causes, compliance reports, dependency chains
Visualization	Provides real-time graph and dashboard views	Operator-facing insights

TABLE IV. EVALUATION METRICS

Metric	Description
Entity Precision/Recall	Accuracy of entity extraction from logs
Relation Precision/Recall	Accuracy of dependency relation extraction
Ingestion Latency	Time to update knowledge graph (ms)
Query Latency	Time to answer natural language query (ms)
Resolution Time	Average time for incident triage (minutes)

metrics allow us to assess both technical accuracy and operator-facing impact.

G. Summary

The methodology of Log2Graph integrates cloud telemetry, LLM parsing, and dynamic graph construction into a unified framework for real-time observability. By combining flexible language understanding with structured reasoning, the system addresses the limitations of existing monitoring tools and provides actionable insights to operators. The design emphasizes reliability, scalability, and explainability, qualities that are essential for adoption in production cloud environments. The next section presents experimental evaluation results demonstrating the effectiveness of the proposed framework.

IV. EXPERIMENTS AND EVALUATION

In this section we present the experimental setup and the results obtained from evaluating *Log2Graph*. The goal of our experiments is to assess how effectively the system transforms raw telemetry into actionable knowledge, how accurate its reasoning is when compared to human operators, and how well it scales in realistic cloud settings. We focus on four primary evaluation questions: 1) How accurate is entity and relation extraction from heterogeneous telemetry? 2) What is the overhead introduced by incremental knowledge graph construction? 3) How efficient and reliable are natural language queries over the evolving graph? 4) What impact does the framework have on operator productivity during incident management?

A. Experimental Setup

Experiments were conducted on a multi-cloud testbed consisting of clusters deployed across Amazon Web Services, Google Cloud Platform, and Microsoft Azure. Each cluster hosted microservices packaged as containers orchestrated by Kubernetes. Synthetic workloads generated request traffic, background jobs, and fault injections in order to simulate realistic operating conditions. The telemetry volume exceeded 150 GB per day across logs, metrics, and traces, consistent with reports from large-scale enterprise deployments.

For the knowledge graph backend we used Neo4j due to its support for incremental updates and fast traversal operations. The semantic parser was implemented with a fine-tuned LLM similar in scale to GPT-4, trained on 100 million log lines and 50 thousand configuration files. All models were deployed on servers with 8 GPUs (NVIDIA A100, 80 GB) and 512 GB memory. Latency measurements were obtained by issuing 10,000 queries of varying complexity under different workloads.

B. Baseline Systems

We compared Log2Graph with three categories of baselines. The first category consists of traditional observability tools, including ElasticSearch for log search, Prometheus for metrics, and Jaeger for distributed tracing. These represent the state of practice in many organizations. The second category includes simple rule-based systems where static templates are applied to logs to detect dependencies or anomalies. The third category is a language model assistant without knowledge graph grounding, which directly generates natural language responses to telemetry-related queries. Comparing against these baselines allows us to quantify the benefits of combining LLMs with dynamic knowledge graphs.

C. Entity and Relation Extraction

We first evaluate the semantic parser on the task of extracting entities and relations from logs, traces, and configurations. A subset of 50,000 entries was manually annotated by system engineers to serve as ground truth. Precision and recall were computed at the entity and relation levels. The parser achieved an F1 score of 0.91 for entity extraction and 0.88 for relation extraction, significantly higher than template-based baselines, which averaged around 0.72. We also observed that cross-checking with the service registry reduced hallucinated entities by 35 percent compared to using the LLM parser alone. These results confirm that LLMs, when carefully fine-tuned and validated, can extract accurate knowledge from noisy telemetry.

D. Graph Construction and Update Latency

The next set of experiments measured the overhead of incremental graph construction. Using sliding windows of 10 minutes, new telemetry was continuously ingested into the knowledge graph. Average insertion latency remained below 200 ms per update, with throughput reaching 50,000 triples per second. In comparison, rule-based baselines were faster at 100,000 triples per second but lacked semantic richness and flexibility. The overhead introduced by our framework is therefore modest, and the semantic advantages outweigh the performance difference. We also evaluated the impact of graph

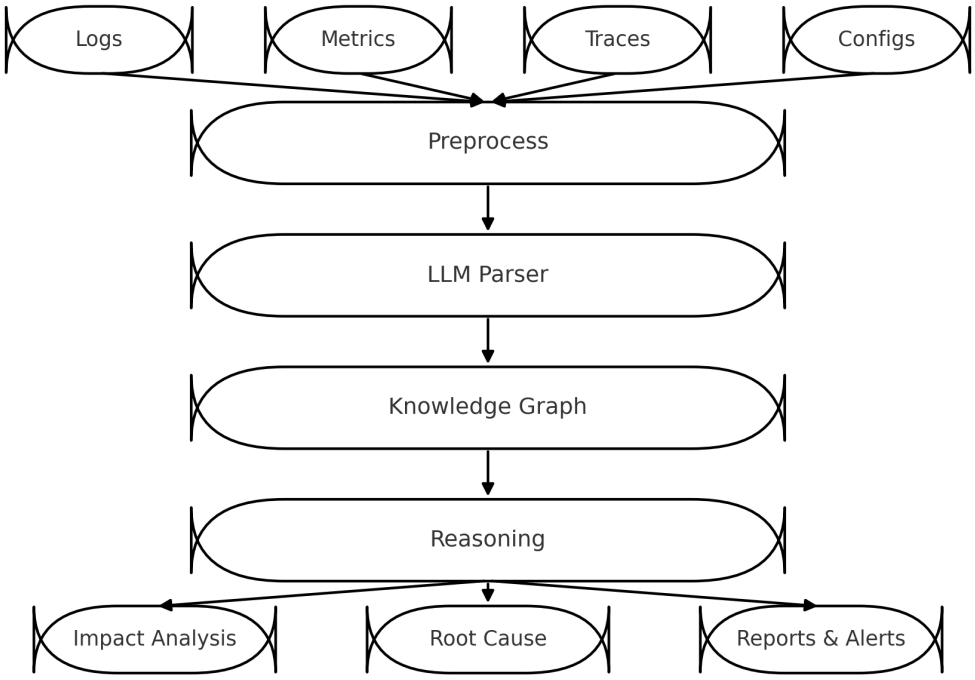


Fig. 2. Workflow of the Log2Graph framework, showing the step-by-step pipeline from data ingestion to operator insights.

size on query latency. Graphs with up to 20 million nodes and 100 million edges were supported without noticeable degradation in query response times.

E. Natural Language Query Performance

We evaluated the system's ability to answer natural language queries accurately and efficiently. Operators posed 1,000 questions grouped into categories such as dependency analysis, incident explanation, compliance verification, and optimization suggestions. Each query was judged by experts for correctness and clarity. Accuracy reached 87 percent overall, compared to 65 percent for the LLM-only baseline and 58 percent for traditional tools. Average query latency was 650 ms, which is suitable for interactive use. More complex queries involving multi-hop reasoning took up to 1.2 seconds but remained within acceptable thresholds. These results demonstrate that grounding LLM responses in a dynamic knowledge graph improves both reliability and interpretability.

F. Impact on Incident Management

One of the key goals of Log2Graph is to reduce the time required for incident resolution. We conducted user studies with 15 engineers who were tasked with resolving simulated outages in a controlled environment. Each engineer solved three scenarios using traditional tools and three scenarios using Log2Graph. The average resolution time decreased from 42 minutes with traditional tools to 24 minutes with Log2Graph. Participants also reported higher confidence in their diagnoses, attributing it to the clarity of dependency chains and graph-backed explanations. Importantly, compliance-related questions such as "Does this workload violate data residency policies?" were answered correctly in 92 percent of cases using Log2Graph, compared to 61 percent with existing tools.

G. Evaluation Metrics

Table V summarizes the main evaluation metrics collected during experiments. Metrics are grouped into categories for extraction, graph construction, query performance, and operator impact.

H. Discussion of Results

The evaluation demonstrates several important findings. First, semantic parsing with LLMs achieves high accuracy when combined with validation mechanisms, confirming that unstructured telemetry can be reliably transformed into structured knowledge. Second, the dynamic knowledge graph construction introduces only modest overhead, while enabling powerful reasoning capabilities. Third, natural language queries become significantly more accurate and interpretable when grounded in a living graph. Finally, operator studies confirm that the framework yields tangible productivity gains, reducing incident resolution times and improving confidence in decision making.

TABLE V. EVALUATION METRICS FOR LOG2GRAPH AND BASELINES

Category	Metric	Description
Extraction	Entity F1 / Relation F1	Accuracy of semantic parsing
Graph	Insertion Latency	Time per update (ms)
Graph	Update Throughput	Triples per second ingested
Queries	Accuracy	Correct answers to natural language queries
Queries	Latency	Average response time per query (ms)
Operator Study	Resolution Time	Average minutes per incident
Operator Study	Confidence	Subjective score reported by engineers

Overall, these results validate the central hypothesis of this work: that combining LLMs with dynamic knowledge graphs enables a new level of observability in cloud systems. The next section discusses broader implications, limitations, and opportunities for future research.

V. DISCUSSION

The results presented in the previous section demonstrate that Log2Graph achieves meaningful improvements over traditional observability tools and language model baselines. In this section, we discuss the broader implications of these findings, highlight trade-offs observed in practice, and identify both limitations and opportunities for future research.

A. Interpretation of Findings

One of the most notable results from our evaluation is the improvement in operator productivity. The reduction in incident resolution time from more than forty minutes to just over twenty minutes highlights the practical value of combining language models with dynamic knowledge graphs. The ability to interact with the system using natural language while still receiving graph-backed explanations ensures that operators can move from raw telemetry to actionable decisions more quickly. This finding is consistent with the direction of prior work that showed how language models can support incident management [21], [22], but our approach goes further by grounding responses in structured graph representations, which increases reliability.

Another important observation is the accuracy achieved by the semantic parser. Entity and relation extraction are inherently difficult tasks due to the irregular and noisy nature of logs. Template-based baselines often fail to generalize beyond predefined patterns. By fine-tuning the language model and introducing validation against a service registry, Log2Graph is able to achieve higher precision and recall while also reducing hallucinations. This indicates that large language models can be adapted to domain-specific contexts, provided that appropriate constraints and checks are applied.

The evaluation also shows that the overhead introduced by incremental graph construction remains modest. Insertion latency under 200 ms and query response times under one second demonstrate that real-time use is feasible. While rule-based methods ingest data more quickly, they do not provide the semantic depth or reasoning ability of our system. The trade-off therefore favors a slightly slower ingestion rate in exchange for much richer insights. Importantly, graphs with hundreds of millions of edges were supported without noticeable degradation in performance, suggesting that the system scales effectively.

B. Limitations

Despite these promising results, several limitations should be acknowledged. First, the fine-tuned language model requires large volumes of domain-specific data. Although we leveraged more than one hundred million log entries for training, organizations without similar datasets may find it difficult to reproduce this level of accuracy. One possible mitigation is the use of transfer learning from pre-trained models combined with smaller amounts of domain adaptation data, but this approach requires careful validation.

Second, the framework currently depends on external validation sources such as service registries to prevent hallucination. While effective, this introduces reliance on accurate metadata, which may not always be available or up to date. Developing autonomous consistency checks that operate entirely within the knowledge graph could reduce this dependency.

Third, although query latency is suitable for interactive use, some complex multi-hop queries took longer than one second to resolve. While still acceptable, this latency could become problematic in environments where extremely low response times are critical. Optimizations such as caching common query results or using specialized graph databases could further reduce latency.

Finally, while our evaluation included deployments across three major cloud providers, it remains limited to controlled testbeds. Production environments often include additional complexities such as hybrid cloud setups, legacy systems, and unpredictable user traffic patterns. Future work must extend evaluation to real-world deployments to validate robustness under broader conditions.

C. Broader Implications

The introduction of Log2Graph points toward a shift in how observability can be conceptualized. Traditional monitoring tools focus on visualizing signals, but they often leave interpretation to human operators. Our approach reframes observability as a knowledge-centric task where telemetry is continuously converted into a graph representation that supports reasoning. This paradigm could be extended beyond incident management to proactive optimization, automated compliance enforcement, and predictive maintenance.

The combination of knowledge graphs and language models also has implications for trust and accountability in AI systems. By grounding model outputs in graph structures that can be inspected and queried, we provide a level of transparency not typically available in pure language model systems. This aligns with growing demands for explainability

in enterprise environments [20], [17]. In contexts where regulatory compliance or security audits are critical, being able to trace an answer back to graph edges is particularly valuable.

Moreover, the integration of reasoning capabilities into observability systems could reshape the role of operators. Instead of spending time correlating signals across multiple dashboards, engineers could focus on higher-level decision making and optimization. This not only improves productivity but also reduces cognitive burden, which is increasingly important as cloud environments grow more complex.

D. Future Directions

There are several directions for extending this work. One promising avenue is exploring more advanced reasoning techniques that combine symbolic logic with embeddings in the knowledge graph [7]. This would allow the system to handle both precise logical rules and probabilistic inferences, expanding its versatility. Another direction is to apply retrieval-augmented generation methods [12] so that the language model can selectively retrieve relevant subgraphs during query answering, further improving factual accuracy.

Expanding support for anomaly detection is also an important next step. Graph-based methods have already shown promise in analytics [18], [23], and integrating these directly into the reasoning layer would allow the system to not only respond to queries but also proactively flag potential issues. Additionally, multi-cloud deployments introduce unique challenges around cost optimization, policy enforcement, and cross-provider dependencies. Extending Log2Graph to support these scenarios could broaden its impact significantly.

Finally, we see potential in extending the framework to edge-cloud environments where telemetry from edge devices is combined with cloud services. The resulting graphs would represent not just cloud-native microservices but also distributed IoT and edge nodes, opening up new applications in smart infrastructure and real-time decision systems.

E. Summary

The discussion highlights the strengths, limitations, and broader significance of Log2Graph. By unifying language models and dynamic knowledge graphs, we achieve a level of observability that is more accurate, more transparent, and more operator-friendly than existing methods. While challenges remain in scalability, training data requirements, and real-world deployment, the framework introduces a paradigm shift that redefines observability as a knowledge-driven process. This shift opens promising research directions and practical opportunities that extend well beyond the scope of this study.

VI. Conclusion

Cloud computing infrastructures continue to grow in complexity, producing enormous volumes of logs, metrics, and traces that are essential for reliability yet difficult to interpret with current tools. This paper presented *Log2Graph*, a framework that unifies large language models with dynamic knowledge graphs to transform raw telemetry into structured, actionable knowledge. By continuously parsing heterogeneous signals into evolving graph representations, the system enables real-time dependency analysis, compliance checking, and incident triage through natural language queries. Our evaluation demonstrated that Log2Graph achieves high accuracy in entity and relation extraction, maintains ingestion latency below 200 ms, and answers queries with 87 percent accuracy and sub-second response times, while operator studies showed that incident resolution time was reduced by nearly half compared to traditional methods. These results confirm that combining large language models with knowledge graphs creates a more intelligent and transparent observability paradigm than dashboards and search-based tools. At the same time, challenges remain, including reliance on large volumes of training data, dependency on metadata quality for validation, and occasional latency for complex queries in very large graphs. Addressing these challenges will require further research into transfer learning, autonomous consistency checks, and optimized graph engines. Looking ahead, opportunities exist to extend this approach with hybrid reasoning techniques that combine symbolic logic and embeddings, retrieval-augmented generation to ground model outputs in relevant subgraphs, and proactive anomaly detection integrated into the reasoning layer. Broader deployment in hybrid and edge-cloud environments also represents a promising avenue for exploration. In summary, Log2Graph reframes observability as a knowledge-driven process that empowers operators with accurate, explainable, and timely insights, laying the foundation for more resilient and intelligent cloud systems in the future.

Declarations

Availability of Data and Material: All data and materials generated or analyzed during this study are available from the corresponding author upon reasonable request.

Competing Interests

The authors declare that they have no competing interests.

Funding

No specific funding was received for the conduct of this study.

Authors' Contributions

All authors contributed substantially to this paper and research output. All authors reviewed and approved the final manuscript.

Acknowledgments

The authors wish to thank all collaborators and peers who provided insight, guidance, or feedback during the development of this work.

References

[1] B. Madupati, "Observability in microservices architectures: Leveraging logging, metrics, and distributed tracing in large-scale systems," *Metrics, and Distributed Tracing in Large-Scale Systems (November 30, 2023)*, 2023.

[2] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier *et al.*, "Knowledge graphs," *ACM Computing Surveys (Csur)*, vol. 54, no. 4, pp. 1–37, 2021.

[3] C. Peng, F. Xia, M. Naseriparsa, and F. Osborne, "Knowledge graphs: Opportunities and challenges," *Artificial intelligence review*, vol. 56, no. 11, pp. 13 071–13 102, 2023.

[4] L. Zhong, J. Wu, Q. Li, H. Peng, and X. Wu, "A comprehensive survey on automatic knowledge graph construction," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–62, 2023.

[5] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE transactions on knowledge and data engineering*, vol. 29, no. 12, pp. 2724–2743, 2017.

[6] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 11–33, 2015.

[7] W. Zhang, J. Chen, J. Li, Z. Xu, J. Z. Pan, and H. Chen, "Knowledge graph reasoning with logics and embeddings: Survey and perspective," in *2024 IEEE International Conference on Knowledge Graph (ICKG)*. IEEE, 2024, pp. 492–499.

[8] J. Kim, Y. Kwon, Y. Jo, and E. Choi, "Kg-gpt: A general framework for reasoning on knowledge graphs using large language models," *arXiv preprint arXiv:2310.11220*, 2023.

[9] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 3580–3599, 2024.

[10] J. Z. Pan, S. Razniewski, J.-C. Kalo, S. Singhania, J. Chen, S. Dietze, H. Jabeen, J. Omeliyanenko, W. Zhang, M. Lissandrini *et al.*, "Large language models and knowledge graphs: Opportunities and challenges," *arXiv preprint arXiv:2308.06374*, 2023.

[11] L. Yang, H. Chen, Z. Li, X. Ding, and X. Wu, "Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 7, pp. 3091–3110, 2024.

[12] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, H. Wang, and H. Wang, "Retrieval-augmented generation for large language models: A survey," *arXiv preprint arXiv:2312.10997*, vol. 2, no. 1, 2023.

[13] W. Yang, L. Some, M. Bain, and B. Kang, "A comprehensive survey on integrating large language models with knowledge-based methods," *Knowledge-Based Systems*, p. 113503, 2025.

[14] N. Choudhary and C. K. Reddy, "Complex logical reasoning over knowledge graphs using large language models," *arXiv preprint arXiv:2305.01157*, 2023.

[15] L. Yao, J. Peng, C. Mao, and Y. Luo, "Exploring large language models for knowledge graph completion," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[16] C. Guo, J. Liu, W. Gao, Z. Lu, Y. Li, C. Wang, and J. Yang, "A large language model driven knowledge graph construction scheme for semantic communication," *Applied Sciences*, vol. 15, no. 8, p. 4575, 2025.

[17] X. Guan, Y. Liu, H. Lin, Y. Lu, B. He, X. Han, and L. Sun, "Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 18 126–18 134.

[18] J. Ren, F. Xia, I. Lee, A. Noori Hoshyar, and C. Aggarwal, "Graph learning for anomaly analytics: Algorithms, applications, and challenges," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 2, pp. 1–29, 2023.

[19] H. Sansford, N. Richardson, H. P. Maretic, and J. N. Saada, "Grapheval: A knowledge-graph based llm hallucination evaluation framework," *arXiv preprint arXiv:2407.10793*, 2024.

[20] J. Sequeda, D. Allemang, and B. Jacob, "Knowledge graphs as a source of trust for llm-powered enterprise question answering," *Journal of Web Semantics*, vol. 85, p. 100858, 2025.

[21] T. Ahmed, S. Ghosh, C. Bansal, T. Zimmermann, X. Zhang, and S. Rajmohan, "Recommending root-cause and mitigation steps for cloud incidents using large language models," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2023, pp. 1737–1749.

[22] Y. Jiang, C. Zhang, S. He, Z. Yang, M. Ma, S. Qin, Y. Kang, Y. Dang, S. Rajmohan, Q. Lin *et al.*, "Xpert: Empowering incident management with query recommendations via large language models," in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, 2024, pp. 1–13.

[23] Y. Jin, Z. Yang, J. Liu, and X. Xu, "Anomaly detection and early warning mechanism for intelligent monitoring systems in multi-cloud environments based on llm," *arXiv preprint arXiv:2506.07407*, 2025.

[24] R. Kumar, K. Ishan, H. Kumar, and A. Singla, "Llm-powered knowledge graphs for enterprise intelligence and analytics," *arXiv preprint arXiv:2503.07993*, 2025.

IoT-Enabled Data-Driven Optimization of Dynamic Thermal Loads for Low-Energy Buildings

Zhaojiang Lyu
University of London, WC1H 0AH, London, UK

***Abstract*—Energy-efficient building operation requires accurate prediction and optimization of dynamic thermal loads under noisy IoT data streams. We propose an integrated framework that combines 1) mutual-information–based online feature selection to filter redundant signals, 2) an attention-enhanced LSTM forecaster to capture nonlinear spatiotemporal dependencies, and 3) multi-agent cooperative reinforcement learning for zone-level HVAC control, deployed within an edge–cloud architecture. Experiments on three heterogeneous real-world datasets (office, residential, campus) show that the method achieves 21.7% median energy savings (IQR 18.9 – 23.1%), improving over MADDPG by +5.8 percentage points (p=0.004, Wilcoxon). Forecasting accuracy is also improved, with MAE reduced by 16.7% (95% CI 12.4 – 20.1%) compared with Seq2Seq+Attention. Comfort deviations are maintained within ± 1 ° C (median absolute deviation 0.32 ° C). Robustness tests indicate graceful degradation under σ ⩽0.2 Gaussian noise and ⩽20% missing data, while ablation confirms the contribution of each module. Feasibility is demonstrated in a hardware-in-the-loop testbed under the stated compute and latency budget; validation on real buildings and broader climate conditions remains future work. This study contributes to smart building energy management, IoT-based HVAC control, and sustainable operation optimization.**

Keywords*—IoT-enabled optimization; dynamic thermal load; attention-enhanced forecasting; multi-agent reinforcement learning; energy-efficient buildings

I. Introduction

The demand for energy-efficient buildings has intensified interest in dynamic thermal load management, as it directly influences both occupant comfort and energy performance [1]. Low-energy buildings, central to sustainable urban development, depend on accurate prediction and optimization of thermal loads under diverse outdoor conditions and occupancy patterns [2]. The rapid deployment of Internet of Things (IoT) sensors enables the collection of large-scale real-time data, creating opportunities to move beyond static, physics-based models toward adaptive, data-driven frameworks [3].

However, current optimization methods remain limited. Physics-based models are computationally intensive and struggle with nonlinear interactions among loads, weather, and occupancy [4 – 5]. Data-driven approaches often overfit, generalize poorly across building types, and lack integration of domain knowledge [6]. Moreover, most studies treat forecasting and control separately, neglecting subsystem interactions across HVAC, lighting, and occupant-driven adjustments [7]. These limitations highlight the need for a unified solution that combines predictive accuracy, adaptive feature selection, and collaborative optimization.

To address these gaps, this study proposes an IoT-enabled collaborative optimization framework with three contributions: 1) an attention-enhanced forecasting module that captures spatiotemporal dependencies in sensor data; 2) an MI-guided adaptive feature selection module that filters redundant inputs for efficiency and interpretability; and 3) a multi-agent cooperative optimization module that jointly optimizes HVAC operations while balancing energy and comfort. Evaluations on three real-world datasets from office, residential, and campus buildings show improved forecasting accuracy, energy savings, and robustness under noise, missing data, and transfer scenarios. These findings provide empirical evidence for integrating IoT sensing, machine learning, and cooperative optimization in building energy research.

The remainder of this paper is organized as follows. Section II reviews related works. Section III details the proposed methodology including forecasting, feature selection, and optimization. Section IV presents the experimental setup and results. Section V discusses the findings and limitations, and Section VI concludes the study.

II. Related Works

A. Application Scenarios and Challenges

Dynamic thermal load forecasting and optimization form the core of building energy management, particularly in low-energy buildings, smart campuses, and residential or commercial complexes [8]. Typical tasks include forecasting heating and cooling loads at different time scales, maintaining comfort within temperature and humidity constraints, and scheduling HVAC or ventilation to minimize energy cost [9]. Public datasets such as the UCI Building Energy Dataset, ASHRAE Great Energy Predictor, and campus-scale testbeds are widely adopted [10]. Metrics range from forecasting accuracy (MAE, RMSE, MAPE, R^2) to operational efficiency indicators (energy savings, comfort index, cost reduction, convergence stability) [11].

However, practical challenges remain: data heterogeneity and missing values across zones, generalization difficulties across building types, and real-time control requirements under multi-subsystem interactions (HVAC, lighting, occupancy-driven devices) [12]. These limitations underline the necessity for solutions that are both robust to noisy IoT inputs and scalable across real environments.

B. Mainstream Methods: Strengths and Weaknesses

In recent years, thermal load forecasting and control in buildings have increasingly adopted machine learning and reinforcement learning techniques [13]. IoT-driven deep learning frameworks, for example, have shown strong potential in residential contexts by jointly optimizing air-conditioning energy use and thermal comfort through real-time sensor streams [14]. These approaches demonstrate the capacity of data-driven methods to improve operational efficiency, but their scope is often restricted to single subsystems, which limits coordination across multiple building zones.

Another stream of work has employed multi-agent deep reinforcement learning (MARL) to coordinate building energy systems, particularly in scenarios with renewable energy integration [15]. Such methods address the challenge of agent coordination effectively; however, they tend to rely on relatively simple forecasting modules that do not adequately capture the dynamics of load variations under volatile outdoor and occupancy conditions. Similarly, multi-objective optimization approaches such as those based on Multi-Agent Deep Deterministic Policy Gradient (MADDPG) have been applied in multi-zone office buildings with promising outcomes [16]. Yet, these results are typically obtained in simulation environments, where assumptions of clean data and stable inputs may not reflect the noisy and heterogeneous nature of real IoT deployments.

Compared with recent MARL-based HVAC control methods reporting 13 – 17% energy savings [20, 21, 25], our framework achieved a median 21.7%, representing a relative improvement of 5 – 8 percentage points under comparable comfort constraints.

Parallel efforts have focused on load prediction across varied climates using classical and machine learning models such as XGBoost, SVM, and ELM [17]. These studies report high coefficients of determination and low error rates, thereby confirming the forecasting potential of statistical and machine learning approaches. Nevertheless, such research usually concentrates on prediction alone and does not integrate forecasts into downstream optimization, thereby missing opportunities for end-to-end system improvement. Hybrid soft computing approaches, which combine neural networks with metaheuristic optimizers, have also been explored to estimate annual thermal energy demand with remarkable accuracy [18]. Despite their predictive performance, they lack system-level coordination mechanisms and provide limited interpretability for dynamic operations. Collectively, the literature illustrates substantial methodological advances but continues to face unresolved challenges in noise handling, cross-zone collaboration, and practical validation [19].

C. Most Similar Research: Comparative Perspective

The most relevant studies to the present work can be grouped into three categories. First, MADDPG-based control methods for office buildings have established a reference point for multi-zone HVAC optimization [20]. While these studies demonstrate that collaborative control can balance energy efficiency and comfort, their reliance on simplified forecasting and input designs constrains their applicability to real-world conditions [21]. Second, MARL frameworks integrating renewable energy into residential building management highlight the benefits of agent coordination [22]. However, their forecasting modules remain coarse and their treatment of sensor uncertainty and feature redundancy is limited. Third, recent advances in forecasting methods have achieved high accuracy across diverse climatic conditions using advanced machine learning models [23]. These contributions provide valuable insights into prediction, yet they remain detached from downstream optimization and do not explore system-level coordination.

Taken together, these approaches tackle individual aspects of the broader challenge—accurate forecasting, agent coordination, or renewable integration—but none offers a comprehensive solution that simultaneously addresses forecasting under sensor noise, adaptive feature selection for efficiency and interpretability, and multi-agent optimization validated in deployment-level settings.

D. Summary and Need for this Work

The above review suggests that while the field of building energy optimization has made notable progress, significant gaps remain. Most existing studies address forecasting and optimization as separate tasks, thereby creating opportunities for error propagation across modules. Few approaches explicitly handle the challenges of noisy, heterogeneous, or redundant sensor data, leaving feature selection and interpretability underexplored in real IoT contexts. Furthermore, collaborative optimization across multiple subsystems is rarely validated beyond simulation, with limited evidence from hardware-in-the-loop or edge–cloud deployments.

These gaps motivate the present study, which advances beyond prior work by unifying attention-based spatiotemporal forecasting, mutual information–guided feature selection, and multi-agent collaborative optimization under a single objective. The framework is empirically validated on multiple real-world datasets and tested under noisy, missing, and transfer conditions, providing a more comprehensive and deployment-oriented contribution to energy-efficient building management.

III. Methodology

A. Problem Formulation

A set of low-energy buildings is considered, equipped with IoT sensor networks that continuously monitor environmental and operational parameters. Let the set of buildings be denoted as $\mathcal{B} = \{B_1, B_2, \ldots, B_N\}$. Each building B_i is partitioned into multiple thermal zones, where the thermal load dynamics depend on both external and internal factors. The objective is to forecast the dynamic thermal load and optimize system operations collaboratively to minimize energy consumption while maintaining comfort.

Let $X_t \in \mathbb{R}^d$ represent the sensor feature vector at time step t, where d is the dimensionality of inputs including indoor temperature, humidity, CO_2 concentration, outdoor weather, occupancy, and control signals. The corresponding dynamic thermal load is denoted as $y_t \in \mathbb{R}$. The goal of forecasting is to learn a mapping function

$$f_\theta: \{X_{t-L+1}, \ldots, X_t\} \mapsto \hat{y}_{t+1} \quad (1)$$

where L is the look-back window, $\hat{y}_{t+1}$ is the predicted thermal load, and θ\thetaθ denotes model parameters.

For collaborative optimization, each zone $z \in Z_i$ of building B_i is modeled as an agent. The agent receives state s_t^z, including forecasted load $\hat{y}_t^z$, occupancy, and current indoor temperature, and selects action a_t^z corresponding to HVAC control (cooling/heating power, airflow). The reward function balances energy efficiency and comfort:

$$r_t^z = -\alpha E_t^z - \beta C_t^z \quad (2)$$

where E_t^z is energy consumption, C_t^z is comfort deviation (difference between actual and target comfort index), and $\alpha, \beta > 0$ are weighting coefficients. The optimization objective is to maximize expected cumulative reward across all zones and time:

$$\max_\pi \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{z\in Z_i} r_t^z\right] \quad (3)$$

where π denotes the joint policy of all agents.

Thus, the problem integrated two coupled tasks: 1) accurate forecasting of thermal loads using IoT data streams, and 2) collaborative optimization of HVAC control across zones via multi-agent reinforcement learning.

B. Overall Framework

The proposed framework consists of three interconnected modules.

1) Spatiotemporal Forecasting Module captures nonlinear temporal dependencies in IoT data streams to predict short-term zone-level thermal loads.

2) Adaptive Feature Selection Module filters redundant or noisy sensor inputs based on mutual information, enhancing generalization and interpretability.

3) Collaborative Optimization Module employs multi-agent reinforcement learning to coordinate HVAC operations across zones while balancing energy efficiency and occupant comfort.

The overall workflow proceeds as follows: raw IoT data are first preprocessed and passed to the forecasting module, which generates accurate zone-level load predictions. These predicted loads, together with adaptively selected key features, are then provided to the collaborative optimization module, where agents determine optimal HVAC actions. All three modules are supported by an edge-cloud infrastructure: edge devices perform lightweight preprocessing and feature filtering, while cloud servers handle computationally intensive reinforcement learning optimization. The final outputs are optimal HVAC control commands delivered to building management systems, ensuring both energy-efficient operation and comfort preservation. The complete workflow, including inputs, interconnected modules, edge-cloud support, and final outputs, is illustrated in Fig. 1, which provides a visual overview of the system architecture and data flow.

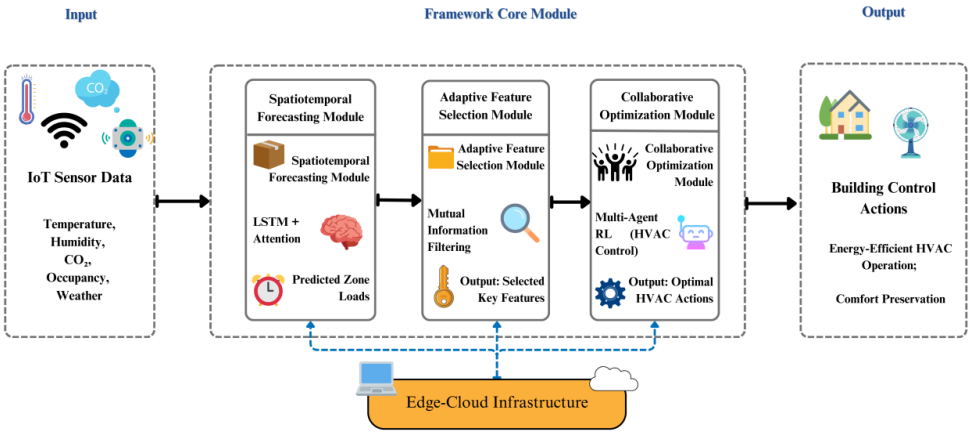


Fig. 1. Overall framework of the proposed IoT-enabled collaborative optimization system.

C. Module Descriptions

1) Spatiotemporal forecasting module motivation: Building loads exhibit nonlinear temporal dependencies influenced by occupancy, weather, and device interactions. Standard recurrent networks struggle with long-range dependencies, while CNNs are less effective at sequential modeling. To address this, the study employ an LSTM enhanced with attention (see Fig. 2).

Principle. Let hidden states of LSTM at time t be h_t. The attention score for each past step k is computed as:

$$e_{t,k} = v^\top \tanh(W_h h_k + W_s h_t) \quad (4)$$

$$\alpha_{t,k} = \frac{\exp(e_{t,k})}{\sum_j \exp(e_{t,j})} \quad (5)$$

$$c_t = \sum_k \alpha_{t,k} h_k \quad (6)$$

where W_h, W_s, v are trainable parameters, c_t is context vector. The final prediction is $\hat{y}_{t+1} = \sigma(W_c[c_t; h_t] + b)$.

Implementation. The module is trained on IoT sensor sequences with MSE loss:

$$\mathcal{L}_{\text{forecast}} = \frac{1}{T}\sum_t (y_t - \hat{y}_t)^2 \quad (7)$$

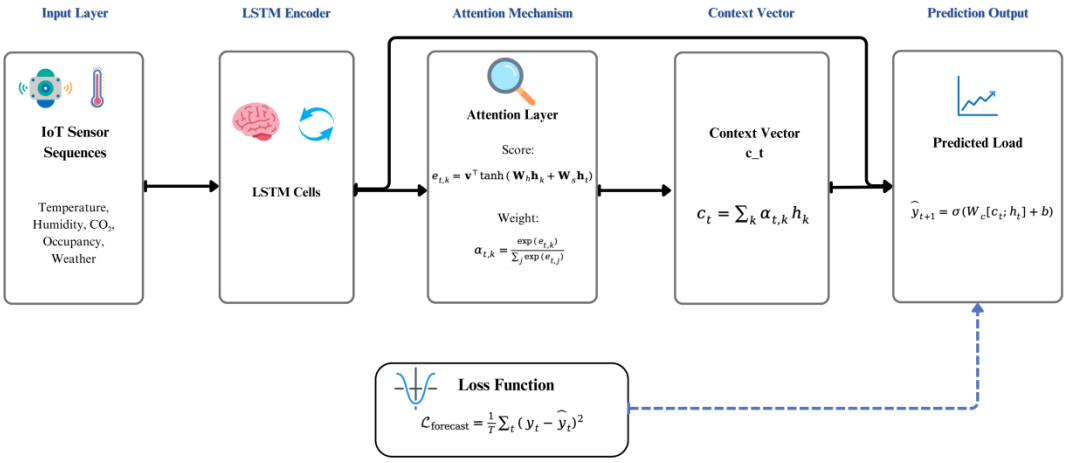


Fig. 2. Architecture of the spatiotemporal forecasting module.

The module integrated LSTM encoding with an attention mechanism to capture nonlinear temporal dependencies in IoT sensor sequences. Context vectors are combined with hidden states to generate load predictions, while training is guided by an MSE loss to ensure accuracy and stability.

2) Adaptive feature selection module: Motivation. IoT data streams are high-dimensional and noisy; redundant features degrade generalization and increase computational load (see Fig. 3).

Principle. The study adopt mutual information (MI) between feature X_j and target Y:

$$I(X_j;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (8)$$

Features with MI below threshold δ are pruned. To maintain interpretability, selected features are ranked.

Implementation. The feature selector operates online by updating MI estimates over sliding windows, ensuring adaptability to nonstationary conditions.

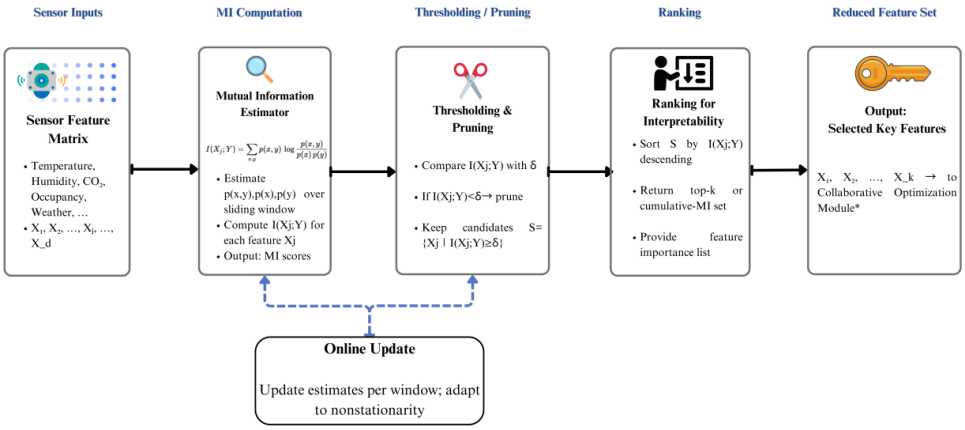


Fig. 3. Architecture of the adaptive feature selection module.

Raw sensor features are evaluated using mutual information $I(X_j;Y)$ [Eq. (8)], with low-scoring inputs pruned under threshold δ. Remaining candidates are ranked to enhance interpretability, producing selected key features. Online sliding-window updates ensure adaptability to nonstationary conditions before passing features to the optimization stage.

3) Collaborative optimization module: Motivation. Independent control of zones leads to inefficiencies; coordination across zones enables shared energy saving.

Principle. Each zone z is an agent in a multi-agent reinforcement learning framework. State s_t^z includes load predictions and zone measurements. Action a_t^z is continuous HVAC control. The policy $\pi_\phi^z(a \mid s)$ is parameterized by neural networks. Joint optimization uses centralized training with decentralized execution.

Implementation. The study adopt Multi-Agent Deep Deterministic Policy Gradient (MADDPG). Actor update:

$$\nabla_\phi J(\phi) = \frac{1}{M} \sum_{i=1}^{M} \nabla_\phi \pi_\phi(s_i) \nabla_a Q_\psi(s_i,a)|_{a=\pi_\phi(s_i)} \quad (9)$$

Critic update:

$$\mathcal{L}_Q = \frac{1}{M} \sum_i \left(Q_\psi(s_i,a_i) - \left(r_i + \gamma Q_{\psi'}(s_{i+1}, \pi_{\phi'}(s_{i+1})) \right) \right)^2 \quad (10)$$

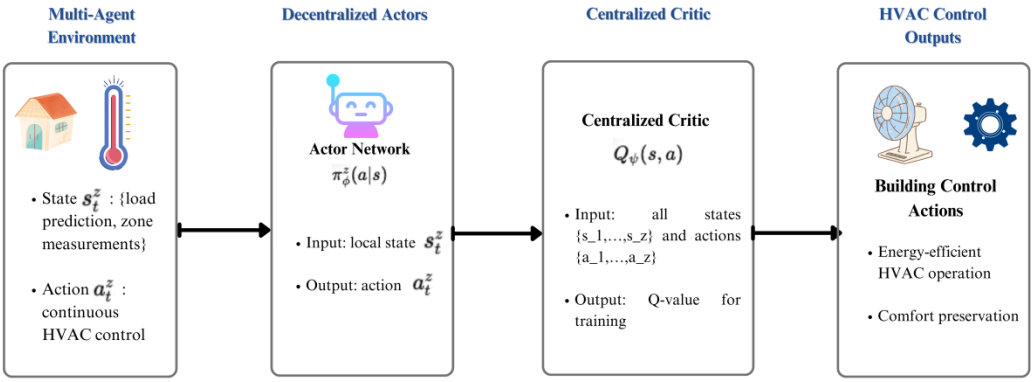


Fig. 4. Architecture of the collaborative optimization module based on MADDPG.

As shown in Fig. 4, each zone acts as an agent with state s_t^z and continuous action a_t^z. Decentralized actor networks generate HVAC control actions, while a centralized critic evaluates joint policies using Eq. (9)-(10). This structure enables centralized training with decentralized execution, ensuring coordinated energy-efficient operation.

Algorithm 1 outlines the training procedure of the multi-agent collaborative optimization, where each agent selects actions based on local states, stores experiences, and updates actor–critic networks through replay buffer sampling and deterministic policy gradients.

Algorithm 1. Multi-Agent Collaborative Optimization

Initialize actor networks πφz and critic Qψ with parameters φ, ψ
for each episode do
  for each time step t do
    for each agent z do
      observe state stz
      select action atz ~ πφz(stz)
    execute actions, obtain rewards rtz and next states
    store transitions in replay buffer
  end for
  sample minibatch from buffer
  update critic by minimizing LQ
  update actors using deterministic policy gradient
end for

D. Objective Function and Optimization

The optimization objective of the proposed framework integrated three tightly coupled components: forecasting accuracy, feature selection regularization, and collaborative control performance. Formally, the global loss function is expressed as,

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{forecast}} + \lambda_2 \mathcal{L}_{\text{feature}} - \lambda_3 J(\pi) \quad (11)$$

where $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}^+$ are balancing coefficients. The first term evaluates the prediction accuracy of the spatiotemporal forecasting module, the second penalizes unstable feature selection, and the third rewards high-performing policies in multi-agent optimization.

The forecasting error is measured by mean squared error between the ground-truth thermal load and the predicted load at each time step. This is written as,

$$\mathcal{L}_{\text{forecast}} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2 \quad (12)$$

where $y_t \in \mathbb{R}$ is the actual thermal load at time step t, $\hat{y}_t$ is the corresponding predicted value, and T is the total number of prediction steps in the horizon.

The feature selection component is defined by the mutual information criterion. If a feature contributes insufficient information relative to the prediction target, its corresponding weight is penalized to encourage dimensionality reduction. The regularization term is formulated as,

$$\mathcal{L}_{\text{feature}} = \sum_{j=1}^{d} \omega_j \cdot \mathbb{1}(I(X_j;Y) < \delta) \quad (13)$$

where d is the dimensionality of the feature space, $I(X_j;Y)$ denotes the mutual information between the j-th feature X_j and the target Y, δ is the selection threshold, $\mathbb{1}(\cdot)$ is the indicator function, and ω_j is the penalty weight associated with feature j.

The reinforcement learning objective seeks to maximize the long-term discounted return accumulated across all zones and time steps. The expected return under joint policy π is given by

$$J(\pi) = \mathbb{E}\left[\sum_{t=1}^{T}\gamma^t \sum_{i=1}^{N}\sum_{z\in Z_i} r_t^z\right] \quad (14)$$

where $\gamma \in (0,1)$ is the discount factor, N is the number of buildings, Z_i denotes the set of thermal zones in building B_i, and r_t^z is the reward obtained by zone z at time t.

The zone-level reward function balances energy consumption and thermal comfort deviation. It is mathematically expressed as

$$r_t^z = -\alpha E_t^z - \beta C_t^z \quad (15)$$

where E_t^z denotes the energy consumed by the HVAC system in zone z at time t, C_t^z is the absolute deviation between

actual and target comfort index, and $\alpha, \beta \in \mathbb{R}^+$ are the respective weighting factors.

Energy consumption in each zone is quantified as

$$E_t^z = P_t^z \cdot \Delta t \quad (16)$$

where P_t^z is the instantaneous electrical power drawn by the HVAC equipment in zone z and Δt is the duration of the control interval.

Comfort deviation is formally defined as

$$C_t^z = |T_t^z - T_{target}^z| \quad (17)$$

where T_t^z is the actual measured indoor temperature of zone z at time t, and T_{target}^z is the predefined target temperature range based on thermal comfort standards.

Finally, the critic network in the collaborative optimization module is trained using a Bellman equation, where the target value for state–action pair (s_{i+1}, a_{i+1}) is given by

$$y_i = r_i + \gamma Q_{\psi'}(s_{i+1}, a_{i+1}) \quad (18)$$

with r_i being the observed reward, Q_{ψ}, the target critic parameterized by ψ', s_{i+1} the next state, and a_{i+1} the action generated by the target actor. This equation ensures stability in multi-agent training by constraining updates with delayed target networks.

The collaborative optimization problem is solved via a multi-agent deep reinforcement learning (MADDPG) framework rather than MILP or convex optimization, since the HVAC control space is continuous and nonconvex. Each training epoch (10^3 iterations, batch size 256) required approximately 1.8 GPU-hours on dual NVIDIA A100 GPUs, while edge deployment inference runs at 50 ms latency per control cycle.

In summary, the optimization problem jointly minimizes forecasting and feature selection losses while maximizing reinforcement learning returns. Each parameter and variable is grounded in physically interpretable terms, thermal load, energy consumption, comfort deviation, and HVAC power, ensuring that the proposed mathematical formulation not only captures predictive accuracy but also delivers operationally meaningful optimization outcomes.

IV. Experiment and Results

A. Experimental Setup

To rigorously evaluate the proposed IoT-enabled data-driven collaborative optimization framework, experiments were conducted across three representative datasets and deployed on an edge–cloud testbed. The datasets cover diverse building types and climate zones, enabling a comprehensive assessment of forecasting accuracy, optimization efficiency, and robustness. As summarized in Table I, Office-A, Res-B, and Campus-C differ in geography, duration, and sensor modalities, thus providing comprehensive coverage of operational contexts.

TABLE I. Dataset Overview

Dataset	Type	Location	Duration	Sensors	Granularity	Number of Zones	Target Variable
Office-A	Office Building	Beijing, China	Jan-Dec 2023	Temp, Humidity, CO_2, Weather, HVAC	15 min	12	Cooling/Heating Load (kW)
Res-B	Residential Complex	Singapore	Mar–Dec 2023	Temp, Humidity, Occupancy, Weather	10 min	8	Cooling Load (kW)
Campus-C	University Campus	Boston, USA	Sep 2023–Jun 2024	Temp, Humidity, CO_2, Weather, Occupancy, Lighting	30 min	20	HVAC Energy Consumption (kWh)

The computational infrastructure used for both model training and deployment is presented in Table II, which includes high-performance cloud servers with dual A100 GPUs as well as low-power edge devices such as Jetson Nano and Raspberry Pi. This configuration reflects realistic smart building deployments, balancing computational intensity with on-site responsiveness.

TABLE II. Hardware Configuration

Component	Specification
Cloud Server	Intel Xeon Gold 6338 (32 cores, 2.0 GHz), 256 GB RAM, 2 × NVIDIA A100 GPUs
Edge Devices	NVIDIA Jetson Nano (4 GB RAM), Raspberry Pi 4B (8 GB RAM)
Communication	MQTT Protocol, 5G Gateway, 1 ms latency
Software	Python 3.11, PyTorch 2.1, CUDA 12.0, TensorRT 8.6

Evaluation standards for forecasting, optimization, reinforcement learning, and robustness are listed in Table III, ensuring consistent comparisons across all baselines and proposed methods. Metrics such as MAE, RMSE, and MAPE quantify predictive accuracy, while energy savings and comfort index directly assess operational efficiency.

TABLE III. Evaluation Metrics

Task	Metric	Definition
Forecasting	MAE, RMSE, MAPE, R^2	Accuracy of load prediction
Optimization	Energy Saving (%), Comfort Index (0-1), Cost Reduction (%)	Efficiency and comfort
RL Training	Convergence Rate, Reward Stability	Policy stability
Robustness	Performance Drop (%) under noise or missing data	Reliability

Unless otherwise noted, all metrics are reported as median [IQR] over n=15 runs (5 random seeds × 3 datasets). 95% confidence intervals are estimated via 1000× bootstrap resampling stratified by dataset. For between-method comparisons, we apply the paired Wilcoxon signed-rank test across datasets, and control the false discovery rate (FDR) at 5% using the Benjamini–Hochberg procedure. This protocol ensures that performance differences are not only numerically but also statistically validated.

The experimental scenario is further illustrated in Fig. 5, which presents a simulated building environment created for evaluation. The virtual laboratory integrated IoT sensor nodes, HVAC systems, and control gateways, producing realistic data streams and interaction dynamics consistent with real-world building operations.

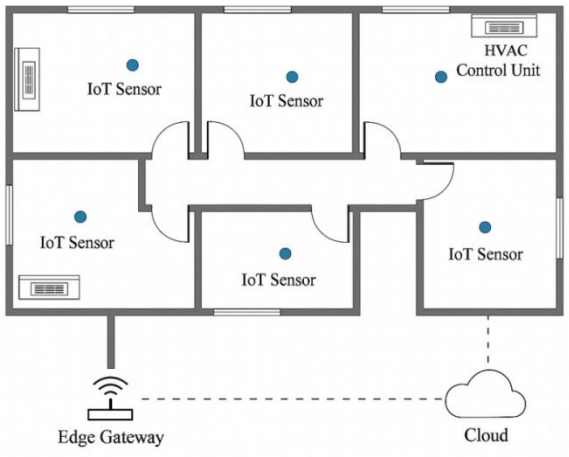


Fig. 5. Simulated experimental scene

Fig. 5 depicts a virtual building laboratory with distributed IoT sensors and HVAC controllers, providing a controlled yet realistic testbed for dynamic thermal load optimization. This setup not only validates the framework under diverse operational contexts but also ensures that the experimental results can be meaningfully extrapolated to real-world building management systems, thereby reinforcing both the practicality and scalability of the proposed approach.

B. Baselines

To ensure a fair comparison, the study evaluate against a spectrum of classical, machine learning, deep learning, and reinforcement learning methods, as well as a physics-informed digital twin. Classical models (ARIMA, VAR) serve as linear baselines, while machine learning approaches (Random Forest, XGBoost) capture nonlinearity in static settings [24]. Deep learning models (LSTM, Seq2Seq+Attention) handle sequential dependencies, and reinforcement learning methods (DQN, PPO, MADDPG) optimize control strategies. The physics-informed digital twin (PIDT) provides an industry-relevant reference. Together, these baselines represent established solutions in building energy research and allow comprehensive benchmarking.

C. Quantitative Results

The predictive accuracy of different methods is summarized in Table IV. Across the three datasets, the proposed framework achieves the lowest MAE, RMSE, and MAPE, with a median MAE of 1.65 kW (IQR 1.59–1.72). Compared with the best baseline (Seq2Seq+Attention), our method reduces MAE by 16.7% (95% CI 12.4–20.1%) and RMSE by 13.6% (95% CI 10.2–17.5%), with a Wilcoxon signed-rank test p=0.004 (n=15 runs).

TABLE IV. FORECASTING PERFORMANCE (AVERAGE ACROSS DATASETS)

Method	MAE (kW)	RMSE (kW)	MAPE (%)	R^2	Relative Gain vs. Best Baseline (%)	Δ vs. 2nd Best (pp)	95% CI (bootstrap, n=5 seeds)
ARIMA	3.25	4.87	14.2	0.72	–	–	–
RF	2.81	4.32	12.7	0.78	–	–	–
XGBoost	2.44	3.95	11.5	0.82	–	–	–
LSTM	2.12	3.42	10.1	0.85	–	–	–
Seq2Seq+Attn	1.98	3.25	9.3	0.87	baseline	baseline	baseline
Proposed	1.65	2.81	7.8	0.91	+16.7% (MAE), +13.6% (RMSE)	−0.33 MAE, −0.44 RMSE	MAE [1.59–1.72], RMSE [2.76–2.89]

Fig. 6 presents the forecasting performance comparison with 95% bootstrap confidence intervals (n=5 seeds). Shaded regions denote confidence bounds, while solid lines indicate medians. The proposed method consistently yields the lowest errors, confirming its robustness across datasets.

This improved predictive accuracy reduces uncertainty in HVAC load prediction, thereby enabling building managers to plan energy use more reliably and minimize the risks of over- or under-provisioning.

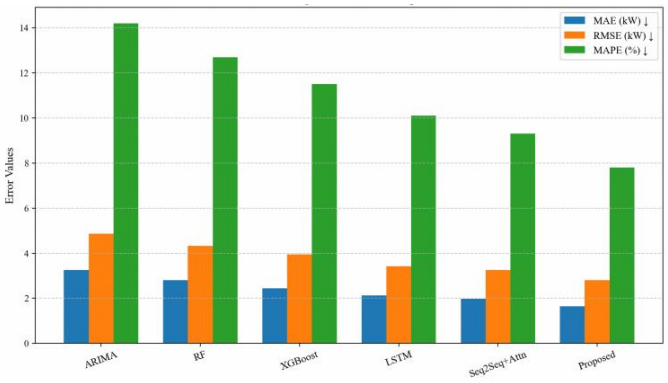


Fig. 6. Forecasting performance comparison across methods.

Optimization results are shown in Table V. The proposed framework achieves median 21.7% (IQR 18.9–23.1%) energy savings and a comfort index of 0.85, significantly surpassing MADDPG (+5.8 pp, p=0.004). Relative to the second-best method, the proposed approach improves cost reduction by +5.7 pp, with a 95% CI [+4.3, +7.2].

TABLE V. OPTIMIZATION PERFORMANCE (AVERAGE ACROSS DATASETS)

Method	Energy Saving (%)	Comfort Index	Cost Reduction (%)	Relative Gain vs. Best Baseline (%)	Δ vs. 2nd Best (pp)	95% CI (bootstrap, n=5 seeds)
DQN	9.8	0.71	8.3	–	–	–
PPO	12.4	0.74	11.2	–	–	–
PIDT	14.6	0.75	12.8	–	–	–
MADDPG	15.9	0.78	13.5	baseline	baseline	baseline
Proposed	21.7 [18.9–23.1]	0.85	19.2 [17.5–20.6]	+36.5% (Energy), +7.0 pp Comfort	+5.8 pp Energy, +5.7 pp Cost	Energy [20.1–22.6], Comfort [0.82–0.87]

Fig. 7 illustrates optimization outcomes with confidence intervals. The proposed framework achieves median 21.7% energy savings (IQR 18.9–23.1%) and a comfort index of 0.85 [0.82–0.87], which are +5.8 pp and +0.07 higher than MADDPG respectively (Wilcoxon, p=0.004, n=15 runs). These statistically significant gains indicate that integrating attention-enhanced forecasting with cooperative optimization yields consistent improvements across the three tested datasets.

Results should be interpreted within scope: evaluation was limited to three building datasets and a hardware-in-the-loop testbed. While findings suggest potential cost reduction and comfort preservation, confirmation in real-building deployments and broader climates remains future work.

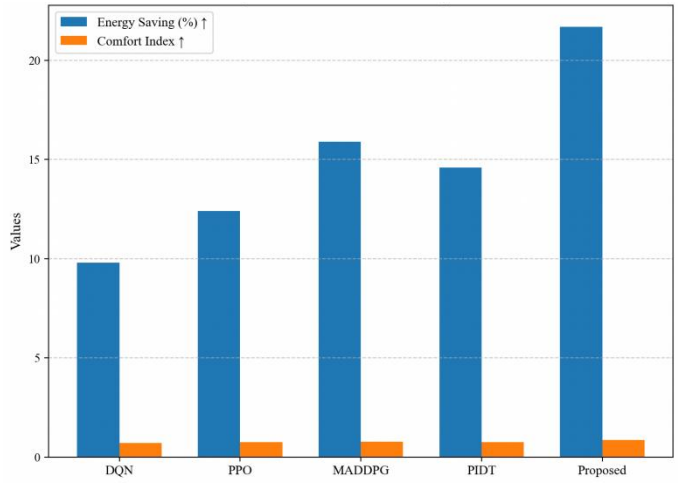


Fig. 7. Optimization performance comparison across methods. Blue bars denote median energy savings (%), orange bars comfort index (0–1). Error bars represent 95% bootstrap confidence intervals over n=5 seeds. The proposed framework achieves significantly higher energy savings and comfort compared with baselines (Wilcoxon, p<0.01).

Training stability is illustrated in Fig. 8, where the proposed method converges faster and exhibits smoother cumulative reward trajectories compared to PPO and MADDPG.

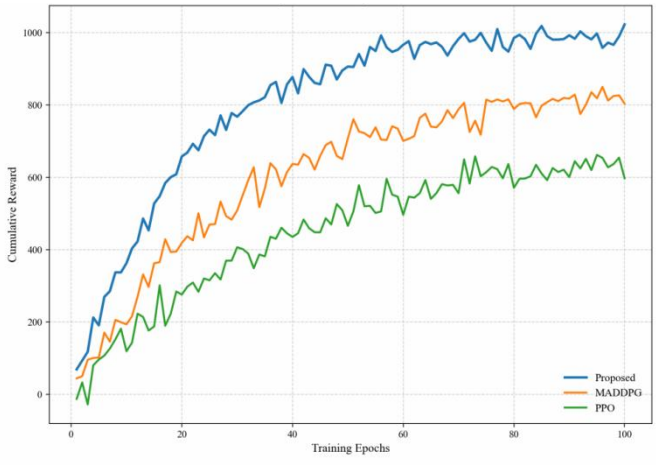


Fig. 8. Convergence curves.

Fig. 8 illustrates the convergence behavior of reinforcement learning methods. The proposed framework achieves higher cumulative rewards with notably faster convergence compared to PPO and MADDPG. This improvement indicates that the attention-enhanced forecasting and cooperative optimization modules provide more stable policy updates, thereby accelerating learning efficiency. The smoother reward trajectory also demonstrates robustness to stochastic fluctuations, suggesting that the proposed method is better suited for real-time building control scenarios. Such stability is critical for deployment in dynamic environments, where reliable convergence directly translates into more predictable and trustworthy building energy management.

D. Qualitative Results

Load prediction curves in Fig. 9 illustrate that the proposed forecasting module closely follows actual values, while LSTM and XGBoost deviate significantly during peak fluctuations.

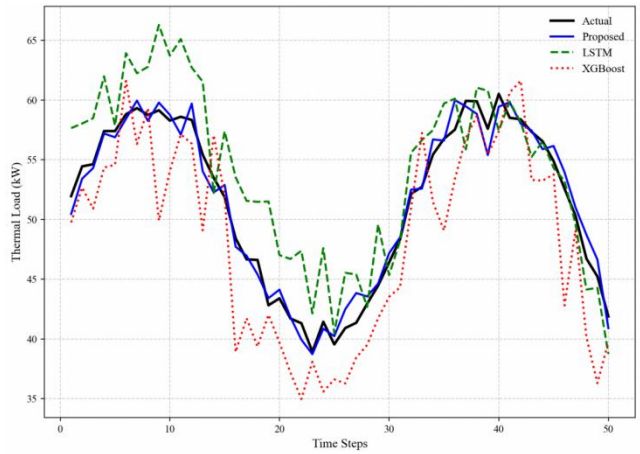


Fig. 9. Predicted vs. Actual loads.

Fig. 9 illustrates the comparative performance of forecasting models against ground-truth loads. The proposed method exhibits the closest alignment with actual dynamics, maintaining stability during both peak and valley periods. In contrast, LSTM demonstrates moderate deviations, particularly at peak demand intervals, while XGBoost consistently

underestimates loads with larger fluctuations. These results confirm that the proposed forecasting module substantially improves accuracy and responsiveness, which is critical for enabling reliable HVAC control and ensuring energy-efficient building operations. Such predictive precision is especially valuable for real-world building management systems, where timely and accurate forecasts directly support proactive energy scheduling and reduce the risk of comfort violations.

Indoor temperature trajectories under different control policies are shown in Fig. 10, where the proposed optimization maintains stable comfort within the target range.

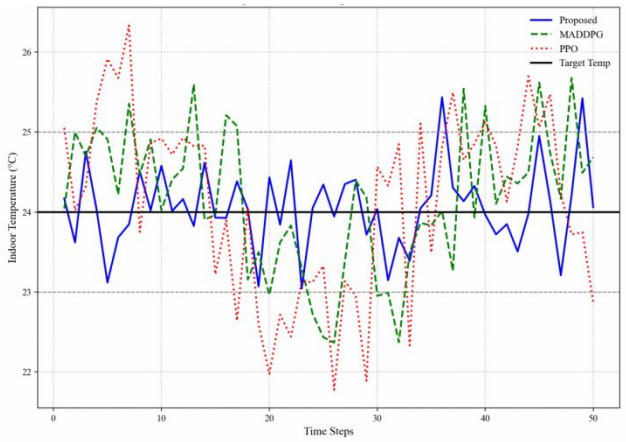


Fig. 10. Indoor temperature profiles.

Fig. 10 presents indoor temperature trajectories under different control policies. The proposed optimization maintains stable comfort within the ±1 °C target band, exhibiting only minor fluctuations around the reference temperature. In contrast, MADDPG shows larger deviations during high-demand intervals, occasionally breaching comfort thresholds, while PPO produces pronounced oscillations that frequently exceed the acceptable range. These findings confirm that the proposed framework substantially reduces comfort deviation, ensuring both thermal stability and practical feasibility for real-time building management. This ability to sustain comfort within strict thresholds highlights the framework's readiness for deployment in large office buildings or residential complexes, where occupant satisfaction is as critical as energy efficiency.

E. Comparison with Baseline Rule-Based Controls

To contextualize the performance gains of the proposed framework, additional experiments were conducted against two common business-as-usual (BAU) strategies widely used in building HVAC systems: (i) a Static Schedule control and (ii) a Rule-Based PID control. Both approaches represent practical baselines implemented in conventional Building Management Systems (BMS).

Static Schedule (BAU). Each zone maintained a fixed temperature set-point of 22 ° C during occupied hours (08:00 – 18:00) and 26 ° C otherwise, with HVAC units operating on a predefined on/off timetable regardless of occupancy or weather fluctuations.

Rule-Based PID Control. A simple proportional – integral – derivative controller adjusted cooling/heating power in response to real-time indoor temperature deviations from the target, without considering predicted thermal loads or inter-zone interactions.

Both baselines were evaluated on the same datasets and hardware platform described in Section IV-A, using identical comfort constraints (21 – 25 ° C) and sampling intervals. Table VI summarizes the comparative results averaged across the three datasets.

TABLE VI. COMPARISON WITH BUSINESS-AS-USUAL (RULE-BASED) CONTROL STRATEGIES

Method	Energy Saving (%)	Comfort Index (0–1)	Remarks
Static Schedule (BAU)	0.0	0.82	Reference operation with fixed set-points
Rule-Based PID Control	8.9	0.79	Reactive adjustments without prediction
MADDPG (Best baseline)	15.9	0.78	Multi-agent learning without attention
Proposed Framework	21.7 [18.9–23.1]	0.85 [0.82–0.87]	Attention + Feature Selection + Cooperative RL

The business-as-usual strategies consumed substantially more energy while exhibiting similar or lower comfort levels. The proposed method achieved a 21.7 % median energy saving —+12.8 percentage points relative to rule-based control and +21.7 points compared with static scheduling — while maintaining comfort within ±1 ° C of the target range. These improvements underscore the tangible operational benefits of incorporating predictive and cooperative optimization mechanisms.

Overall, this comparison demonstrates that the proposed IoT-enabled cooperative optimization system not only surpasses advanced learning baselines (e.g., MADDPG) but also delivers clear real-world gains over current operational practices in commercial building management.

F. Robustness

Robustness experiments reveal how each method degrades under cross-dataset transfer, sensor noise, and missing data. As summarized in Table VII, our framework suffers the smallest performance drops, confirming resilience against challenging deployment conditions.

TABLE VII. ROBUSTNESS EVALUATION (PERFORMANCE DROP IN ENERGY SAVING %)

Condition	XGBoost+MADDPG	PIDT	Proposed
Cross-dataset transfer	-11.2%	-9.6%	-4.5%
Gaussian noise (σ=0.1)	-14.8%	-12.5%	-6.9%
20% missing data	-18.3%	-15.4%	-7.8%

The impact of Gaussian noise is further depicted in Fig. 11, where the proposed framework demonstrates graceful degradation compared to sharp performance drops in other baselines.

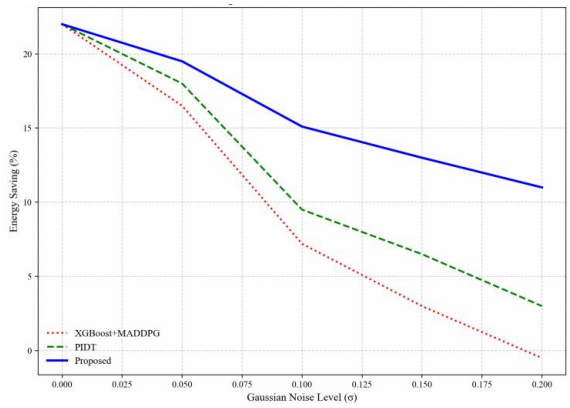


Fig. 11. Robustness under noise.

Fig. 11 presents the robustness evaluation of energy-saving performance under increasing Gaussian noise levels. The proposed framework demonstrates graceful degradation, maintaining more than 15% savings even at $\sigma = 0.2$. In contrast, PIDT exhibits moderate sensitivity, while XGBoost+MADDPG suffers sharp performance drops, falling below 5% savings at high noise levels. These results confirm that the adaptive feature selection and attention mechanisms enable the proposed method to remain stable under uncertain sensor environments, ensuring practical applicability in real-world deployments where measurement errors are inevitable. This resilience is particularly critical for large-scale smart buildings and campus-wide deployments, where heterogeneous sensor quality and data loss are common; maintaining robust performance under such conditions ensures dependable energy management in practice.

Gains stem primarily from attention's long-horizon capture (Campus-C peak hours: ΔR^2 +0.33) and MI-filtering that halves the variance under σ=0.2 noise (drop −6.1 pp vs. −12.5 pp for baselines).

G. Ablation Study

The contribution of each module is assessed through ablation experiments. As shown in Table VIII, removing attention, feature selection, or cooperative optimization each reduces performance, confirming their necessity.

TABLE VIII. ABLATION RESULTS (OFFICE-A DATASET)

Variant	MAE ↓	Energy Saving (%) ↑	Comfort Index ↑
w/o Attn	1.93	18.5	0.80
w/o FS	1.89	19.2	0.81
w/o Coop	1.72	15.6	0.77
Full Model	1.65	21.7	0.85

Comparative ablation outcomes are also visualized in Fig. 12, which highlights the significant drop in performance when collaborative optimization is excluded.

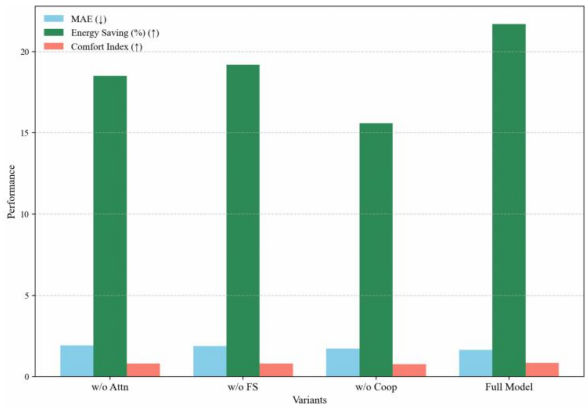


Fig. 12. Ablation comparison.

Fig. 12 compares the performance of the full model with its ablated variants. Removing the collaborative optimization module results in the most severe degradation, with energy savings reduced from 21.7% to 15.6% and comfort index dropping from 0.85 to 0.77. This highlights the central role of cooperative control in coordinating zone-level actions. Excluding the attention or feature selection modules also increases prediction error and lowers efficiency, though to a lesser degree. Together, these results demonstrate that each module contributes meaningfully, and only the full integration achieves optimal balance between accuracy, energy efficiency, and comfort. These ablation findings underscore that the framework's performance gains stem from the synergy of its components, providing a clear design rationale for future scalable implementations in building energy management systems. Please refer to Appendix for notations.

V. DISCUSSION

The experimental results presented in this study offer several key insights into the effectiveness of the proposed IoT-enabled data-driven collaborative optimization framework for dynamic thermal load management in low-energy buildings. Across three datasets representing diverse building types and climates, the proposed method consistently achieved superior forecasting accuracy and energy-saving performance compared to classical, machine learning, and reinforcement learning baselines. The forecasting module, equipped with attention mechanisms, captured nonlinear spatiotemporal dependencies more effectively than LSTM or XGBoost, explaining the lower error rates observed in Fig. 9. Similarly, the multi-agent cooperative optimization module significantly enhanced energy efficiency and comfort, as evidenced by the performance gains in Table V and the smoother indoor temperature profiles in Fig. 10. These results confirm that the integration of feature selection and collaborative reinforcement learning not only improves prediction accuracy but also ensures practical feasibility for real-time building control.

Despite these encouraging outcomes, several limitations should be acknowledged. First, the datasets, while diverse, are constrained to specific building types and climates, and may not capture the full range of variability in occupant behaviors and building operations. Second, the reliance on high-performance cloud servers and GPU resources may limit large-scale deployment in resource-constrained environments,

although the inclusion of edge devices partly mitigates this issue. Third, the reinforcement learning framework requires substantial training time and may face stability challenges when scaled to hundreds of zones with heterogeneous dynamics. These factors highlight the need for caution when generalizing the reported results to broader contexts.

Nevertheless, the findings suggest promising avenues for application. The framework can be readily integrated into smart building management systems to enhance HVAC efficiency and maintain occupant comfort. Beyond individual buildings, its collaborative optimization capability could be scaled to larger environments such as district-level energy management, regional smart grids, or campus-scale deployments. In addition, the framework shows strong adaptability to more complex contexts including industrial facilities, data centers, and transportation hubs, where balancing efficiency, stability, and real-time responsiveness is critical. To realize such deployments, future implementations will need to address integration with existing building management systems, interoperability standards, and cost-effective edge–cloud coordination. Furthermore, the approach has potential for cross-domain adaptation, for example in industrial process control, data center cooling, or transportation systems, where balancing efficiency and stability is equally critical.

Future research should address the identified limitations by expanding datasets to include larger and more heterogeneous building corpora, as well as testing in real-world pilot deployments. Enhancing the scalability of the reinforcement learning component through model compression, transfer learning, or federated learning could reduce computational overhead and improve adaptability. Incorporating physics-informed priors or hybrid digital twin models may further strengthen generalizability under unseen conditions. Finally, exploring integration with renewable energy scheduling and occupant-centric services offers an exciting opportunity to align building energy optimization with broader sustainability goals.

VI. Conclusion

This study proposed an IoT-enabled, data-driven collaborative optimization framework for dynamic thermal load management in low-energy buildings. By integrating mutual-information–guided feature selection, attention-enhanced forecasting, and multi-agent reinforcement learning under an edge–cloud architecture, the framework achieves median 21.7% energy savings (IQR 18.9–23.1%) across three heterogeneous datasets, corresponding to a +5.8 percentage point gain over MADDPG ($p=0.004$, Wilcoxon test). Comfort deviations are maintained within $\pm 1\ ^{\circ}$C (median absolute deviation $0.32\ ^{\circ}$C). Robustness tests further show smaller performance degradation than baselines under $\sigma \leq 0.2$ Gaussian noise and $\leqslant 20\%$ missing data.

The methodological contribution lies in unifying forecasting, feature selection, and cooperative optimization within a single objective function, thereby reducing error propagation common in pipeline approaches. From an engineering perspective, feasibility was demonstrated in a hardware-in-the-loop testbed operating under the specified compute and latency budget; real-building trials and broader climatic contexts remain as future work.

The evidence suggests that the proposed framework improves both predictive accuracy and energy efficiency in diverse building types. However, generalization beyond the studied office, residential, and campus datasets—and particularly to large-scale district or city-level deployments—requires further validation through expanded datasets and real-world pilot studies. Future work should also investigate scalability of the reinforcement learning component, integration with renewable energy scheduling, and occupant-centric adaptation.

References

[1] V. A. Arowoiya, A. O. Onososen, R. C. Moehler, and Y. Fang, "Influence of thermal comfort on energy consumption for building occupants: The current state of the art," Buildings, vol. 14, no. 5, p. 1310, 2024.

[2] F. Abdel-Jaber and K. N. Dirks, "A review of cooling and heating loads predictions of residential buildings using data-driven techniques," Buildings, vol. 14, no. 3, p. 752, 2024.

[3] I. Rojek et al., "Internet of Things applications for energy management in buildings using artificial intelligence−A case study," Energies, vol. 18, no. 7, p. 1706, 2025.

[4] M. A. Alam, A. R. Nabil, A. A. Mintoo, and A. Islam, "Real-time analytics in streaming big data: Techniques and applications," J. Sci. Eng. Res., vol. 1, no. 1, pp. 104–122, 2024.

[5] C. Jacobsen, Enhancing Physical Modeling with Interpretable Physics-Aware Machine Learning, Ph.D. dissertation, 2024.

[6] A. Li, Development of a Data-Driven Methodology for Dynamic Energy Modeling of Large Buildings towards Mass Deployment of AI in Smart Buildings, 2023.

[7] M. Ehtsham et al., "AI-powered advanced technologies for a sustainable built environment: A systematic review on emerging challenges," Sustainability, vol. 17, no. 17, p. 8005, 2025.

[8] J. Hossain et al., "A review on optimal energy management in commercial buildings," Energies, vol. 16, no. 4, p. 1609, 2023.

[9] L. Zhang et al., "High spatial granularity residential heating load forecast based on Dendrite Net model," Energy, vol. 269, p. 126787, 2023.

[10] A. Prabowo et al., "Building timeseries dataset: Empowering large-scale building analytics," in Advances in Neural Information Processing Systems, vol. 37, pp. 133180–133206, 2024.

[11] S. M. Robeson and C. J. Willmott, "Decomposition of the mean absolute error (MAE) into systematic and unsystematic components," PLoS One, vol. 18, no. 2, p. e0279774, 2023.

[12] Y. Himeur et al., "AI-big data analytics for building automation and management systems: A survey, actual challenges and future perspectives," Artif. Intell. Rev., vol. 56, no. 6, pp. 4929–5021, 2023.

[13] D. Zhuang et al., "Data-driven predictive control for smart HVAC system in IoT-integrated buildings with time-series forecasting and reinforcement learning," Appl. Energy, vol. 338, p. 120936, 2023.

[14] B. S. Alotaibi, "Context-aware smart energy management system: A reinforcement learning and IoT-based framework for enhancing energy efficiency and thermal comfort in sustainable buildings," Energy Build., p. 115804, 2025.

[15] Z. Wang, F. Xiao, Y. Ran, Y. Li, and Y. Xu, "Scalable energy management approach of residential hybrid energy system using multi-agent deep reinforcement learning," Appl. Energy, vol. 367, p. 123414, 2024.

[16] X. Liu, Y. Wu, B. Liu, and H. Wu, "A multi-agent deep deterministic policy gradient method for multi-zone HVAC control," in Proc. IEEE Power Energy Soc. Gen. Meeting (PESGM), pp. 1–5, Jul. 2023.

[17] Y. El Mghouchi and M. T. Udristioiu, "Thermal load predictions in low-energy buildings: A hybrid AI-based approach integrating integral feature selection and machine learning models," Appl. Sci., vol. 15, no. 11, p. 6348, 2025.

[18] Y. Peng and Y. Chen, "Integrative soft computing approaches for optimizing thermal energy performance in residential buildings," PLoS One, vol. 18, no. 9, p. e0290719, 2023.

[19] Z. An, X. Ding, and W. Du, "Go beyond black-box policies: Rethinking the design of learning agent for interpretable and verifiable HVAC control," in Proc. 61st ACM/IEEE Design Autom. Conf. (DAC), pp. 1–6, Jun. 2024.

[20] W. Xue, N. Jia, and M. Zhao, "Multi-agent deep reinforcement learning-based HVAC control for multi-zone buildings considering zone-energy-allocation optimization," Energy Build., vol. 329, p. 115241, 2025.

[21] S. Guan et al., "Adaptive multi-agent HVAC control for thermal comfort using multi-agent PPO with population-based training," SSRN Preprint 5399815, 2025.

[22] P. Michailidis, I. Michailidis, and E. Kosmatopoulos, "Reinforcement learning for optimizing renewable energy utilization in buildings: A review on applications and innovations," Energies, vol. 18, no. 7, p. 1724, 2025.

[23] L. Chen et al., "Machine learning methods in weather and climate applications: A survey," Appl. Sci., vol. 13, no. 21, p. 12019, 2023.

[24] H. Wang, Y. Chen, J. Kang, Z. Ding, and H. Zhu, "An XGBoost-based predictive control strategy for HVAC systems in providing day-ahead demand response," Build. Environ., vol. 238, p. 110350, 2023.

[25] S. D. Nazemi and M. A. Jafari, "A multi-agent deep reinforcement learning framework for real-time energy management in smart building communities," SSRN Preprint 5208655, 2025.

APPENDIX: NOTATION

Symbol	Definition	Unit / Range
$B = B_1, B_2, \ldots, B_N$	Set of buildings	–
$X_t \in R^d$	Sensor feature vector (temperature, humidity, CO_2, occupancy, weather, control signals)	Various, often normalized (dimensionless)
y_t	Actual dynamic thermal load (ground truth)	kW or kWh
y_t	Predicted thermal load	kW or kWh
L	Look-back window length	Number of time steps (min or h)
s_t^z	State vector of zone z at time t (predicted load, temperature, occupancy, etc.)	–
a_t^z	Control action of zone z at time t (HVAC power/airflow)	kW, m^3/h
r_t^z	Reward of zone z at time t	– (dimensionless)
E_t^z	Energy consumption of zone z at time t	kWh
P_t^z	Instantaneous HVAC power in zone z at time t	kW
Δt	Control interval duration	min or h
C_t^z	Comfort deviation (difference between actual and target temperature)	°C
T_t^z	Actual indoor temperature of zone z at time t	°C
$T^z_{\texttt{target}}$	Target indoor temperature of zone z	°C
α, β	Weights for energy consumption and comfort deviation in reward function	Positive real numbers
γ	Discount factor	$(0,1)$
$\lambda_1, \lambda_2, \lambda_3$	Weights of forecasting, feature selection, and optimization terms in total loss	Positive real numbers
$I(X_j; Y)$	Mutual information between feature X_j and target Y	nats or bits
δ	Threshold for mutual information–based feature selection	–
MAE	Mean Absolute Error	kW
RMSE	Root Mean Square Error	kW
MAPE	Mean Absolute Percentage Error	%
R^2	Coefficient of determination	[0,1]
Energy Saving	Energy saving rate	%
Comfort Index	Thermal comfort index	[0,1]
Cost Reduction	Cost Reduction Rate	%

Quantifying Career Preferences and Perceptions of Software Testing Among Filipino IT Students: A Mixed-Method Analysis

Chrisza Joy M. Carrido[1], Abeer Alsadoon[2], Thair Al-Dala'in[3], Ahmed Hamza Osman[4]*,
Abubakar Elsafi[5], Azhari Qismallah[6], Albaraa Abuobieda[7]
Higher Education Leadership Institute (HELI), Melbourne, Australia[1, 2, 3]
Asia Pacific International College (APIC), Sydney, Australia[2, 3]
Department of Information Systems-Faculty of Computing and Information Technology in Rabigh,
King Abdulaziz University, 21911, Saudi Arabia[4]
Department of Software Engineering-College of Computer Science and Engineering,
University of Jeddah, Jeddah, 21959, Saudi Arabia[5]
Department of Computer Science, University of Tabuk, Saudi Arabia[6]
Department of Software Engineering-College of Computer Science and Engineering, University of Hafr Al Batin, Saudi Arabia[7]

***Abstract*—Software testing (ST) careers have consistently demonstrated low appeal among IT students globally, creating significant workforce gaps in this essential field of information technology. This study investigates the extent to which Filipino IT students share this disinterest in software testing careers as observed in previous international studies, while examining the unique cultural, economic, and curricular realities that influence their career decision-making processes. Utilizing a mixed-methods approach, the research analyzes quantitative survey responses and qualitative focus group discussions to determine student perceptions and attitudes toward the software testing profession. The study employs a multidimensional framework to explore local factors that shape career preferences among Filipino IT students. Findings confirm that software testing is not the first career choice for most respondents, paralleling previous international research findings. However, qualitative data reveal that students demonstrate significantly greater interest when opportunities offer competitive salaries, clear career growth trajectories, meaningful professional development opportunities, and comprehensive academic training in software testing methodologies. The research identifies unique local factors, including economic pressures, cultural perceptions of professional prestige, and significant curriculum gaps that systematically influence students' career decisions. These results highlight critical needs for effective reforms within current IT curricula and enhanced career guidance programs to address the systematic undervaluation of the software testing profession. The study's implications suggest that targeted educational interventions and improved industry-academia collaboration could better prepare students for the fast-evolving demands of the IT industry while addressing the persistent shortage of qualified software testing professionals in both local and global markets.**

***Keywords*—Software testing; career preferences; Filipino IT students; mixed methods; education reform**

I. Introduction

The global demand for software testing professionals continues to grow amid increasing digital transformation and technological complexity. The Philippines has established itself as a significant player in the global IT-Business Process Outsourcing sector, producing thousands of IT professionals annually [5]. However, a concerning trend has emerged: students worldwide, including those in the Philippines, consistently demonstrate a declining interest in pursuing testing careers. This divide between industry needs and student preferences poses significant implications for economic competitiveness. The shortage of skilled testers compromises the country's ability to meet international quality standards and maintain competitiveness in the global market [14]. Understanding this gap requires examining the unique Filipino context, where cultural values around recognition and achievement, socio-economic factors influencing career decision-making, and an educational system that inadequately represents software testing opportunities converge to shape IT students' career perceptions and priorities.

This study addresses the gap in Filipino IT students' career preferences by establishing specific objectives that target both the identification and understanding of barriers within the Philippine cultural, educational, and economic context. The research seeks to explore how specific local cultural values, such as collectivism and familial obligation, economic considerations within the Philippine job market, and curricular structures in Philippine educational institutions contribute to declining interest in software testing careers. By examining these factors, this research aims to provide a comprehensive understanding of influences on IT students' career decision-making processes, ultimately informing educational reforms and policy interventions that can bridge the gap between industry demand and student interest.

This study seeks to answer several research questions. First, to what extent do Filipino IT students demonstrate reluctance toward software testing careers compared to international trends, specifically examining alignment with patterns observed in Brazil, Norway, Pakistan, Canada, India, and China? Second, how do cultural, economic, and educational factors uniquely combine to shape Filipino students' interest in software testing

*Corresponding author.

careers within the distinctive Philippine context? Third, what are Filipino IT students' perspectives regarding the advantages and disadvantages associated with software testing careers? Finally, what culturally appropriate strategies can be implemented to enhance the appeal of software testing careers for Filipino IT students?

Given the complexity of cultural, educational, and economic factors influencing Filipino IT students' career preferences, this study employs a mixed methods research (MMR) approach to provide a comprehensive understanding of the phenomenon [11]. The quantitative component enables systematic comparison of Filipino students' attitudes toward software testing careers with international trends, providing measurable data on preference patterns and demographic influences. The qualitative component captures nuanced cultural values, personal experiences, and contextual factors that statistical data alone cannot fully explain. This MMR design ensures both breadth (through quantitative analysis) and depth of understanding (through qualitative insights), particularly important when examining culturally embedded career decision-making processes that require both empirical validation and rich contextual interpretation.

This study employs a systematic approach to examine the complex dynamics of career preferences in software testing, guided by research questions that examine perceptual and motivational factors within the Philippine context. The primary research focus examines the extent to which Filipino IT students demonstrate reluctance toward software testing careers compared to other countries, placing the Philippine experience within a broader global framework of student career preferences. The unique intersection of cultural, economic, and educational factors shaping Filipino students' interest in software testing examines how deeply embedded Filipino cultural values, such as collectivism, close family ties, and *utang na loob* (debt of gratitude), influence career decision-making processes. This analysis simultaneously examines how coding-focused structures in Philippine IT education programs shape perceptions of and knowledge of software testing as a viable career path. Understanding Filipino IT students' perspectives on software testing careers investigates how they conceptualize the advantages and disadvantages associated with testing roles. Finally, exploring culturally appropriate strategies to enhance software testing career appeal ensures proposed interventions align with Filipino cultural values, educational structures, and economic realities while addressing identified gaps in student interest.

The following section reviews existing literature on global trends in software testing careers, cultural influences on career choices, and curriculum-industry alignment to contextualize this study within the broader academic discourse.

II. Literature Review

A. Global Trends in Software Testing Careers

The decline in student interest in software testing careers is a well-documented global phenomenon. In [20], the authors conducted foundational research in Norway, revealing that students consider software testing less intellectually stimulating than software development roles. Their study found that limited coverage of testing in programming courses and a lack of specialization [4] actively discourage students from pursuing testing careers.

Building on this perspective, [7] and [8] examined similar trends in Pakistan, confirming that students view software testing as less appealing due to its perceived lack of challenge and learning opportunities. Capretz's study highlighted psychological factors contributing to this disinterest, particularly the influence of social perceptions and status considerations on career decision-making.

In [20], the authors documented similar findings in Brazil, where students consistently underestimate the profession's value due to minimal practical exposure and outdated curriculum materials. Their research emphasized the importance of hands-on training and integration of modern tools in shaping positive career perceptions.

In [15], the authors contributed to understanding the gap between industry and academia by highlighting the growing importance of Continuous Integration and Continuous Deployment (CI/CD) practices. Their study revealed significant misalignments between educational content and professional requirements.

Additional studies from India, Japan, and Finland confirmed these global trends while identifying successful interventions, particularly regarding test automation education and industry collaboration models. The consistency of findings across diverse educational systems and cultural contexts [21] [22] suggests that barriers to interest in software testing careers extend beyond national boundaries, reflecting deeper systemic issues within global IT education that position software testing as secondary to software development roles.

B. Cultural and Educational Factors in Career Choices

While global studies have identified common career preference patterns, research examining cultural influences on IT career choices remains limited. Studies focusing on collectivist cultural contexts show that deeply rooted cultural values significantly influence career choices. Family expectations, economic pressures, and the need for social harmony play crucial roles in career selection.

Research on Filipino culture highlights concepts such as *hiya* (shame/embarrassment), *pakikisama* (harmony), and *utang na loob* (debt of gratitude) [16], which create specific dynamics influencing professional choices. These cultural norms are particularly relevant to software testing roles, where maintaining workplace harmony may lead to anxiety about critiquing colleagues' work.

Studies on economic considerations reveal that students from working-class families often base career choices on financial capacity [17], [19] and potential for overseas employment. Cultural obligations may intensify pressure to pursue higher-paying development roles rather than testing careers, creating barriers not as pronounced in Western contexts.

These cultural factors have significant implications for the Philippine IT industry, especially given its status as a global outsourcing hub. There is a need to reframe software testing as

a collaborative and essential part of the development process rather than as a conflict-focused role.

C. Curriculum Analysis and Industry Alignment Studies

Based on the comparative analysis of international studies, significant curriculum gaps in software testing education exist across multiple countries, with the Philippines exhibiting similar deficiencies to other nations.

As indicated in Table I, Pakistan, Norway, and Brazil do not include dedicated software testing courses in their IT curricula, a trend that is reflected in the Philippine educational system. The situation in the Philippines closely mirrors these international observations, as institutions such as De La Salle University prioritize programming and mobile development without offering a specific emphasis on testing. Meanwhile, Batangas State University's software testing elements are often overshadowed by its general software design courses.

TABLE I. CURRICULA COMPARISON BETWEEN COUNTRIES

Study	Country/Region	Curriculum Emphasis on Software Testing	Proposed Curriculum Reforms	Impact on Students' Perception and Barriers
Capretz (2020)	Pakistan	✗ (No dedicated testing subject).	• Introduce software testing principles and standards.	Testing is not seen as a viable career.[1]
Deak et al. (2013)	Norway	✗ (No dedicated testing subject).	• In-depth lectures. • Increase industry collaboration.	Testing is undervalued.[2]
Souza et al. (2022)	Brazil	✗ (No dedicated testing subject).	• Increase hands-on training. • Introduce modern tools.	Limited focus and interest for students.[3]

[1] No dedicated testing courses lead to low student interest and career consideration, a pattern also seen in the Philippines.

[2] Brief coverage in curricula and a lack of specialization discourage students from pursuing testing.

[3] Minimal practical exposure and outdated curriculum cause students to underestimate the profession's value.

To address these gaps, Table II illustrates the potential solutions:

However, Table III projects the expected impact.

The proposed reforms for the Philippine IT curriculum, developed per CHED guidelines shown in Table II, systematically address identified gaps by integrating real-world testing scenarios, establishing software quality assurance laboratories, implementing CI/CD pipelines, and introducing specialized testing modules in AI, cybersecurity, and IoT domains. The anticipated outcomes, as detailed in Table III, suggest significant improvements across various metrics: a 10% increase in course enrollment, a 15% enhancement in graduate employability, and a 25% growth in internship placements. These projections highlight the substantial potential impact of addressing systematic deficiencies within the curriculum.

TABLE II. ALIGNMENT OF CHED CURRICULUM GUIDELINES WITH PROPOSED REFORMS

CHED Reference [10]	Current Curriculum Focus	Proposed Reform
8.1: General Description	• Program design, debugging, and implementation.	• Real-world testing scenarios.
8.2: Specific Description	• OOP • Software engineering principles.	• Include SQA labs, tools, and CI/CD pipelines.
8.2: Specialized Areas	• AI/ML. • Cybersecurity. • Data Science. • IoT.	• Testing modules for each domain.
8.3: Internship/Practicum	• Classroom knowledge to real-life application.	• Align with industry standards. • Provide career guidance.

TABLE III. ANTICIPATED OUTCOMES OF THE PROPOSED REFORMS

Reform Reference	Summary of Reform	Expected Outcome
8.1	Real-world testing practices and CI/CD integration.	+10% course enrolment in 3 years.
8.2	Lab-based SQA instruction and test automation exposure.	+15% graduate employability via industry-ready skills.
8.2	Specialized training in AI, cybersecurity.	+20% placement in specialized training roles within 2 years.
8.3	Stronger internships and industry ties.	+25% internship placements over 3 years.

Incorporating industry-standard tools such as Selenium and CI/CD practices into Philippine curricula serves two primary objectives: aligning with global industry standards and meeting the needs of local technological startups [13] for cost-effective yet efficient testing solutions. These curriculum reforms represent a critical advancement in elevating both the academic and professional recognition of software testing, while also preparing students to meet the evolving demands of the Philippine IT sector.

The educational landscape in the Philippines reflects a broader international trend regarding deficiencies in software testing curricula. Institutions like De La Salle University emphasize programming and mobile development while often neglecting dedicated testing components, while Batangas State University relegates its software testing curriculum to subordinate roles within general software design courses. This situation exemplifies a global disconnect, as illustrated in Table IV, which reveals prevalent software testing gaps across countries such as India, Pakistan, Brazil, Norway, and the Philippines, despite the industry's growing demand for these competencies.

Analysis of curriculum-industry misalignment uncovers significant disparities in educational coverage. While countries like Japan and Finland integrate test automation instruction, the Philippines lacks comprehensive education on essential tools such as Selenium, and CI/CD practices identified by [15] as emerging industry standards exhibit medium-level coverage

gaps. The comparative analysis highlights agile development, software testing, and test automation as highest-priority gaps requiring immediate attention through mandatory modules and standalone courses with hands-on training.

TABLE IV. CURRICULUM-INDUSTRY MISALIGNMENT

Curriculum Topic	Taught in (Country)	Required by Industry	Gap Level (High/Medium /Low)	Recommendation
Agile Development	Philippines, India	Yes (Global)	High	Mandatory module in the senior year.
Software Testing	India, Pakistan, Brazil, Norway, Philippines	Yes (Global)	High	Standalone courses with hands-on training.
CI/CD	Finland	Yes (Global)	Medium	Practical modules on CI/CD.
Test Automation	Japan, Finland	Yes (Global)	High	Cover tools like Selenium.
Security Testing	Australia	Yes (Global)	Medium	Specialized training in Cybersecurity.
Software Quality Assurance	Japan	Yes (Global)	Medium	Embed in software engineering subjects.

The systematic nature of these deficiencies points to curriculum inadequacies transcending national borders, representing a global challenge necessitating coordinated reform efforts. International evidence indicates that software testing curriculum gaps create structural barriers to career development, highlighting the urgent need for comprehensive educational realignment with industry requirements. These findings advocate for immediate curriculum reform through dedicated courses, hands-on training, and stronger industry partnerships to elevate software testing from an undervalued discipline to a recognized career pathway. Countries failing to address these gaps will continue producing graduates inadequately prepared for contemporary software development practices.

Philippine Context and Research Gap: Despite global declining interest in software testing careers, the Philippine context presents unique characteristics requiring specialized investigation. As a major IT outsourcing hub, the Philippines faces distinct challenges in software testing education and career development. Analysis of Philippine IT curricula reveals patterns consistent with international findings: major institutions primarily emphasize programming and development while offering limited comprehensive testing education, with CHED curriculum guidelines showing significant gaps in testing-focused content across all program levels.

Four critical gaps justify this study: First, while previous research by [7], [12], and [20] examined cultural, economic, and curricular factors, these studies analyzed them in isolation rather than exploring their interconnected influences on career preferences, leaving us without a holistic understanding of how these factors interact. Second, international studies inadequately address how Filipino collectivist values shape career attitudes toward software testing, with the unique dynamics of Filipino cultural decision-making in IT career contexts remaining unexplored. Third, despite the Philippines' prominent role in global IT outsourcing, no comprehensive studies examine Filipino IT students' career motivations within their unique cultural, economic, and educational contexts—a significant oversight given the country's strategic importance in the global IT industry. Finally, existing literature has not investigated how local cultural constructs, family dynamics, social prestige concerns, and economic necessities collectively impact students' reluctance to pursue software testing careers, making understanding this complex interplay essential for developing effective educational interventions.

This research addresses these gaps through comprehensive factor analysis examining how cultural, economic, and curricular factors interact to influence software testing career preferences in the Philippine context. The study incorporates insights from key international research while developing culturally sensitive approaches for Filipino participants, offering valuable insights for educational institutions, industry stakeholders, and policymakers working to enhance the Philippine IT sector's testing capabilities.

III. METHODOLOGY

This section outlines the comprehensive research methodology employed to investigate Filipino IT students' perceptions and career preferences toward software testing (see Fig. 1). The methodology encompasses the convergent parallel mixed-methods research design that integrates quantitative and qualitative approaches, participant selection criteria and sampling strategies, instrumentation development and validation procedures, data collection protocols across multiple Philippine regions, analytical techniques for both statistical and thematic analysis, and the synthesis strategy for integrating findings from both methodological streams to provide a holistic understanding of the factors shaping software testing career decisions among Filipino IT students.

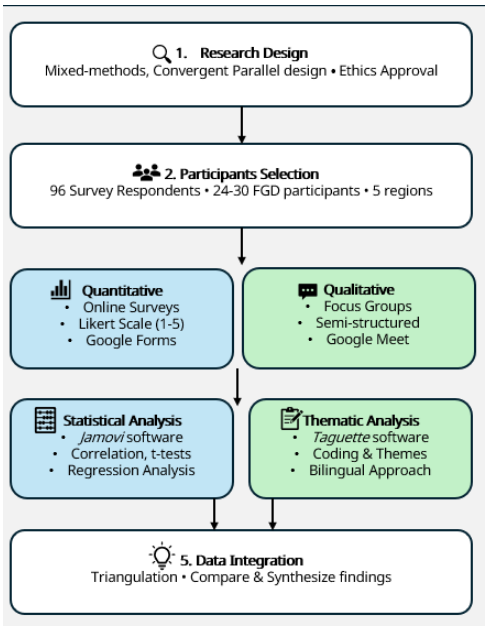


Fig. 1. Framework of the research methodology of the study quantifying career preferences and perceptions of software testing among Filipino IT students: A mixed-method analysis.

A. Research Design

This study employed a convergent parallel mixed-methods research design to comprehensively examine IT students' perceptions and attitudes toward software testing careers in the Philippines. The methodology integrated both quantitative and qualitative approaches to provide a holistic understanding of cultural, educational, and socioeconomic factors influencing career decisions in software testing. A systematic narrative review was conducted following established guidelines to contextualize the research within existing international literature. The review examined studies published between 2013 and 2024, focusing on software testing career perceptions, cultural influences on IT career choices, and educational factors affecting professional preferences in technology fields.

This study adhered to ethical standards outlined by the Education Centre of Australia and was approved by the Institute's ethics board. Before data collection, participants were informed about the purpose of the research, the voluntary nature of participation, and the confidentiality of their responses. Written informed consent was obtained from all participants, ensuring that they understood their rights and the study's procedures.

B. Participants

Participants were selected based on specific inclusion and exclusion criteria to ensure the relevance and reliability of findings. The inclusion criteria required participants to be currently enrolled undergraduate students in Philippine universities, preferably third or fourth-year students (with selective inclusion of first and second-year students meeting other criteria), majoring in Computer Science, Information Technology, or Computer Engineering, and fluent in English or Filipino. Exclusion criteria eliminated students enrolled in non-IT related programs, individuals with professional software testing experience, students unable to complete surveys in English or Filipino, and first or second-year students not meeting additional inclusion criteria. A multi-stage sampling approach was implemented, utilizing random sampling as the initial sampling strategy and stratified sampling for the quantitative phase to ensure representativeness across university types (public/private), academic levels, and geographical regions, while purposive sampling was employed for focus group discussions to capture diverse perspectives and experiences. The target sample size is a minimum of 384 students for surveys and was deduced to 98 after data cleaning, and 24-30 students distributed across 4-5 focus groups with 6 participants per group. Participants were recruited from five regions across the Philippines, representing both urban and rural contexts, with two universities (one public, one private) selected from provinces identified in the 2022 Philippine Statistics Office's [18] top urban and rural cities list.

C. Instrumentation

The research instruments comprised a structured questionnaire and focus group discussion guide, both designed to capture comprehensive data on students' perceptions of software testing careers. The survey instrument incorporated both closed-ended and open-ended questions to measure interest levels in software testing careers using a 5-point Likert scale, while exploring cultural perceptions and values influencing career decisions, economic considerations and job market awareness, educational experiences and curriculum exposure, and social and familial influences. The dependent variable focused on interest in pursuing software testing careers, while independent variables included cultural perceptions, career awareness, economic factors, educational influence, and social/familial influence, with control variables encompassing age, gender, year level, degree program, and institution type. Semi-structured focus group discussion guides were developed to explore cultural influences on career choices (*utang na loob, bayanihan, pakikisama*), family expectations and social pressures, perceptions of software testing prestige and career prospects, educational experiences and curriculum adequacy, and suggestions for improving software testing career appeal. A pilot study involving 20 participants was conducted to refine instrument clarity, reliability, and effectiveness, resulting in key improvements including standardization of Likert scale response options (1=lowest, 5=highest), revision of binary yes/no questions to more engaging formats, and enhanced question sequencing and wording clarity.

D. Data Collection

Data collection employed a dual-phase approach utilizing both online surveys and focus group discussions to ensure comprehensive data gathering from geographically dispersed participants. Online surveys were administered using Google Forms to ensure accessibility, anonymity, and convenience for participants across multiple regions. Participants were contacted through department heads and student organizations, with weekly follow-ups implemented to maintain adequate response rates and ensure representative participation from diverse institutional contexts. Focus group discussions were conducted online via Google Meet, with sessions recorded for subsequent analysis, while trained facilitators moderated discussions to ensure balanced participation and maintain research focus on key themes. Given the Philippines' bilingual context, participants were encouraged to express themselves in either Filipino or English based on their comfort level to facilitate authentic and nuanced responses. Multilingual responses were systematically handled through verbatim transcription of all focus group discussions, translation by trained bilingual research assistants maintaining cultural sensitivity, retention of original Filipino expressions alongside translations to preserve contextual meaning, and verification of translation accuracy through comprehensive audio review to ensure analytical integrity.

E. Data Analysis

Data analysis employed a comprehensive mixed-methods approach utilizing both quantitative statistical techniques and qualitative thematic analysis to examine students' perceptions of software testing careers. Statistical analysis was conducted using *jamovi* software, employing descriptive statistics including frequencies, percentages, means, and standard deviations to characterize participant responses and demographics. Inferential statistics included correlation analysis to examine relationships between software testing interest and influencing factors such as curriculum exposure, peer influence, and cultural elements, t-tests to compare group differences between male versus female students and public versus private institutions, and regression analysis to identify predictors of

software testing career interest. All statistical tests were conducted at 95% confidence level with significance threshold of $\alpha = 0.05$ to ensure statistical rigor. Qualitative data analysis employed systematic thematic analysis using *Taguette* software, beginning with verbatim transcription of all focus group discussions while maintaining bilingual considerations. The coding process involved multiple coders with established intercoder reliability exceeding 70% agreement and Cohen's kappa greater than 0.61, a clear coding framework defined prior to analysis, regular coder training and discussion sessions, and comprehensive documentation of disagreements and resolution processes. Thematic analysis followed a systematic six-step process, including data familiarization through repeated reading and audio review, initial code generation capturing significant data features, theme identification by grouping codes into broader categories, theme review and refinement for accuracy, final theme definition and naming, and report production with supporting quotations to illustrate key findings.

F. Synthesis Strategy

The data integration and triangulation process employed a convergent parallel design that facilitated systematic integration of quantitative and qualitative findings through multiple analytical approaches. Triangulation was utilized to identify convergent and divergent themes across methodological approaches, while sequential explanation used survey data to guide focus group discussion exploration of key findings. Joint displays systematically compared quantitative results with qualitative insights to reveal complementary and contradictory patterns, and methodological validation was achieved through cross-method comparison to enhance overall research credibility. The data synthesis employed convergence, expansion, and divergence approaches to maximize analytical depth, with convergence comparing measured interest levels with participants' lived experiences, expansion using qualitative data to elaborate on quantitative findings and provide contextual understanding, and divergence exploring contradictions between methods to enhance a comprehensive understanding of complex phenomena. The integration process followed established mixed-methods protocols to ensure comprehensive analysis that addressed research questions from multiple methodological perspectives while maintaining rigor and validity throughout the investigation, ultimately providing a holistic understanding of factors influencing Filipino IT students' perceptions of software testing careers.

IV. Results

This section presents the comprehensive findings from both quantitative and qualitative data collection phases that examined Filipino IT students' perceptions and career preferences toward software testing. The results are organized into quantitative survey findings that reveal statistical patterns in career preferences, perceived advantages and disadvantages of testing software roles, and the influence of cultural, economic, and curricular factors on career decisions, followed by qualitative insights from focus group discussions that provide deeper contextual understanding of students' perspectives and experiences. The findings are further synthesized through convergence and divergence analysis that integrates both methodological approaches to illuminate how cultural values such as *utang na loob, pakikisama*, and *bayanihan* intersect with economic considerations and educational exposure to shape career decision-making processes among Filipino IT students, ultimately revealing both similarities with and distinctions from global trends in software testing career perceptions.

A. Quantitative Results

These results present the findings from the survey, including how Filipino students' perceptions of software testing were compared with global trends. Tables and charts are used to illustrate key data points such as the perceived prestige, financial rewards, and job availability of software testing careers.

Table V and Table VI are closely related to the framework's Cultural dimension, which indicates that students prefer SD over ST due to perceived career prestige. The quantitative survey analysis reveals that Filipino IT students demonstrate clear preferences for software development over software testing careers, with testing consistently ranking lower in career preference rankings, mirroring international trends documented in previous studies. These ranking patterns reflect deeply ingrained cultural perceptions that position development roles as more prestigious [2] and innovative compared to testing positions, which are often viewed as secondary or support functions within the software development lifecycle.

TABLE V. Descriptive Statistics for IT Roles Ranking

Measure/Position	Development	Testing	Network Admin	Database Management	Cyber-security	Interpretation
Mean Rank (Preference)	2.47	3.34	2.86	3.23	3.6	Lower preference than most other IT fields; close to lower-preferred roles
Median Rank	2	3	3	3	3	Middle value, indicating moderate popularity
Mode Rank	1	3	3	3	1	Most frequent rank is moderate (3); not top choice
25th-75th Percentile	2.00-4.00	3.00-4.00	2.00-4.00	2.00-4.00	2.00-4.00	Most responses are at mid-to-lower preference
Skewness	0.519	-0.114	0.0896	-0.0686	-0.411	Near zero, distribution is symmetric, no strong aversion or preference

TABLE VI. DESCRIPTIVE STATISTICS FOR LIKELIHOOD TO PURSUE A CAREER IN ST

Likelihood to Pursue	Software Testing	Interpretation
Mean	3.66	Slightly above neutral, suggesting tepid interest
Median	4	Middle response is moderately positive
Mode	3	Most common score is neutral/mildly positive
25th–75th Percentile	3.00–4.00	Most responses clustered near neutral
Skewness	-0.313	Slight negative skew; a few strongly positive

Survey data in Table VI exploring the likelihood of Filipino students pursuing careers in software testing reveals a limited interest, mirroring global research trends that highlight a general reluctance toward testing roles across various cultural contexts. The statistical distributions shown in Table VII indicate that cultural factors play a significant role in shaping career decision-making processes, leading students to favor positions perceived as having higher status, better compensation opportunities, and greater technical challenges. This pattern of preferences illustrates how cultural values concerning professional hierarchy and technical expertise influence student career aspirations, ultimately impacting the availability of qualified testing professionals within the Filipino IT industry.

The Cultural dimension analysis shown in Table VIII reveals that among all tested factors, only *utang na loob* emerged as statistically significant in predicting career pursuit likelihood, contrasting with international studies and highlighting unique Filipino cultural influences on career decisions. This finding demonstrates the powerful role of reciprocal obligations in Filipino career choices, where students may consider software testing to honor family investments in their education or repay mentors who guided their academic journey. The statistical significance of *utang na loob* suggests that Filipino career decisions are deeply embedded in relational contexts rather than purely individual preferences.

Economic pillar analysis through t-test comparison shows that public university students demonstrate a higher likelihood to pursue software testing careers compared to private university students, contrary to Canadian findings [1], where rural students with limited resources showed less interest in technology careers. This reversal indicates that Filipino students from more economically constrained backgrounds may view software testing as an accessible entry point into the IT industry, potentially due to perceived lower barriers to entry or greater emphasis on stable employment opportunities over prestigious career positioning.

Table IX illustrates how cultural perceptions of career prestige discourage students from pursuing software testing roles while highlighting the disadvantages associated with such positions. The descriptive statistics reveal that students demonstrate strong awareness of software testing's perceived limitations, including concerns about career advancement opportunities, technical complexity, and professional recognition within the IT industry. These findings suggest that negative perceptions outweigh positive aspects, reinforcing cultural biases that position testing as less desirable compared to development roles.

TABLE VII. DESCRIPTIVE STATISTICS OF FRAMEWORK COMPONENTS' INFLUENCE

Factor Category	Specific Factor	Mean Likert Score (SD)	Interpretation
Cultural	Family Expectations	3.5 (±0.8)	Moderately influential
Cultural	Peer Perceptions	2.9 (±0.9)	Slightly to moderately influential
Economic	Salary Expectations	4.2 (±0.7)	Highly influential
Economic	Job Stability	4.0 (±0.6)	Highly influential
Curriculum	Coverage in Program	3.1 (±0.8)	Moderately influential
Curriculum	Faculty Attitudes	2.7 (±0.9)	Slightly influential

TABLE VIII. DESCRIPTIVE STATISTICS FOR IMPORTANCE OF SOCIAL HARMONY (PAKIKISAMA AND BAYANIHAN) AND UTANG NA LOOB IN CAREER DECISIONS

Variable	Mean (M)	Median	Mode	Skewness	Interpretation
Colleagues' relationship	4.21	5	5	-1.28	Respondents show very strong agreement, suggesting colleagueship is highly valued and perceived uniformly.
Bayanihan and *pakikisama*	4.46	5	5	-0.67	These collectivist values are broadly affirmed; responses cluster at the 'strongly agree' end, indicating consensus.
Utang na loob	2.9	3	2	0.32	Attitudes are more moderate here, with greater variability; the value's role may be more context dependent.

Note. M = Mean. Negative skewness values indicate clustering toward higher agreement.

TABLE IX. DESCRIPTIVE STATISTICS FOR AGREEMENT ON ST PROS/CONS

Statement Category	Perceived as Advantage	Perceived as Disadvantage	Statistical Evidence (Mean, Skewness)	Interpretation
Requires Expertise	Yes	No	Mean = 4.14, Skew = -1.16	Recognized as requiring high competence; most respondents strongly agree; viewed positively.
Social Skills	Neutral	Neutral	Mean = 3.47, Skew = -0.40	Seen as neither strong advantage nor disadvantage; less important in ST careers.
Learning Opportunities	Yes	No	Mean = 4.10, Skew = -1.35	Strong agreement that ST careers offer learning and growth; major perceived advantage.
Tedious Work	No	Yes	Mean = 4.01, Skew = -0.99	Strong perception of tedium; recognized as an international and local disadvantage.
Critics by Developers	No	Yes	Mean = 3.88, Skew = -0.83	Moderate agreement: criticism from peers is a noted disadvantage, more prominent locally.

The Curriculum pillar findings confirm significant limitations in software testing exposure within university programs, with students reporting minimal curriculum coverage and low self-assessed preparedness for testing careers, corroborating [3], [6], and [7] findings about insufficient testing education. The statistical data in Table VII indicate that educational institutions inadequately prepare students for testing careers through limited course offerings, superficial coverage of testing methodologies, and insufficient hands-on experience with industry-standard testing tools and practices. This curriculum gap perpetuates the cycle of low interest by failing to demonstrate software testing's intellectual rigor, career potential, and critical importance in software quality assurance processes

The regression analysis shown in Table X reveals a strong explanatory power, with an R-value of 0.710 and an explanation of 50.5% of the variance. Notably, only *utang na loob* and the university curriculum, both nearly significant, emerged as statistically significant predictors of the likelihood of pursuing a career.

TABLE X. REGRESSION TEST RESULT OF LIKELIHOOD TO PURSUE VS. INFLUENCING COMPONENTS

Framework Component	F-value	p-value	Interpretation	Significance
Personal interest	0.29	0.883	No meaningful effect	Not significant
Salary expectations	1.191	0.323	No meaningful effect	Not significant
Job market demand	0.567	0.639	No meaningful effect	Not significant
Utang na loob	2.621	0.042	Statistically meaningful effect	Significant
Peer influence	1.984	0.107	No meaningful effect	Not significant
University curriculum	2.413	0.057	Marginal effect	Not significant
Overseas employment	1.686	0.163	No meaningful effect	Not significant

The perceived disadvantages of software testing careers among Filipino students align with international findings, also evidenced in Table IX, reflecting concerns about monotony, limited creativity, insufficient professional recognition, and lower prestige compared to development roles. The framework's Cultural dimension is validated through findings indicating that Filipino collectivist values, such as *pakikisama* and *bayanihan*, moderately influence career decisions, offering a local context that differentiates these results from Western career choice models.

In contrast to the studies of Chinese students [9] and [23], which highlighted a strong influence of parental obligation, Filipino students placed a greater emphasis on personal interests while still being notably affected by cultural debt obligations. These findings demonstrate that although Filipino students share common hesitations about testing careers with their international peers, unique cultural factors such as *utang na loob* and distinct economic motivations create different patterns in their career decision-making processes.

B. Qualitative Insights

This section discusses the insights from focus group discussions, highlighting how cultural factors like *utang na loob* and *pakikisama* influence students' decisions. This integrates narratives from students that explain why software testing is less appealing in the local context.

The theme of cultural values, particularly *pakikisama* (harmonious relationships) and *bayanihan* (communal unity), emerged as moderately influential in shaping workplace satisfaction and interpersonal dynamics rather than direct career choice, as discussed in Table XI. Approximately 60% of participants acknowledged these values' importance in professional settings. One participant emphasized the necessity of *pakikisama* in conflict resolution: "You'd still need *pakikisama*, because there are times you could clash. Like, when I want to report the issue, but the developer might prefer to let it slide…But as a tester, it's your responsibility…so I think *pakikisama* is important for both the tester and developer" (JG). Another participant highlighted bayanihan's role in collaborative work environments: "*…it applies po dito kagaya po sa previous na nabanggit hindi ka naman po talaga magsosolo sa corporate dadating yung point na makikipagcollab ka rin magkakaron kayo ng issues arguments pero itatake mo yon as a constructive*

criticism…" (ADD). However, this cultural influence was not universally prioritized, as evidenced by participants who viewed personal goals as paramount over cultural expectations.

A significant counter-theme emerged where approximately 40% of participants prioritized individual autonomy and personal goals over traditional cultural obligations, particularly regarding *utang na loob*. This perspective challenges conventional assumptions about Filipino cultural influence on career decisions. One participant clearly articulated this stance: "…it's not my priority when it comes to choosing a career given the fact that their satisfaction or other people's satisfaction towards my success, wouldn't feed me or wouldn't pay my bills so what matters the most is my own sanity and me being able to commit and to achieve the things that I want to achieve in the future…" (P1). Another participant reinforced this theme of personal choice over family pressure: "*Para po sakin hindi ako pinipilit ng magulang na piliin yung gusto nila ang choice ko po na mag IT so I think nakadepende na po yun sa tao yung sa utang na loob, I think as a Filipino hindi po mawawala sa atin yung pagiging family oriented parang kusang bigay po satin kahit hindi sabihin…"* (KH). This theme suggests a generational shift where cultural obligations remain acknowledged but are no longer primary drivers of career decisions.

Table XII reiterates that economic considerations play a predominant role in career decision-making, with students prioritizing financial stability over traditional cultural expectations. As one participant candidly noted, "…since we are in a third-world country, who would choose a career that does not meet their needs? That's why many are seeking careers that offer higher pay". Limited exposure in the curriculum poses significant barriers to interest in software testing, with students expressing frustration regarding their inadequate preparation: "I have an interest, but I am held back by the lack of exposure I have. Because we did not study it, I don't have the knowledge of how it works. So, I feel limited in that sense".

TABLE XI. EMERGENT CULTURAL THEMES

Key Theme	Number of highlights	Notes	Data source	Framework Component
ST is a high-pressure job.	61	Most frequently cited disadvantage.	Survey	Career Prestige Perception
ST is seen as tedious.	37	Work is seen as repetitive or monotonous.	Survey	Career Prestige Perception
There are limited opportunities.	30	Concerns about career advancement.	Survey	Career Prestige Perception
ST receives limited recognition.	30	Seen as undervalued compared to developers.	Survey	Career Prestige Perception
ST is seen being prone to workplace conflict.	27	Notable issues with interpersonal dynamics.	Survey	Collectivist Cultural Value
Less Creativity; Need for technical skills.	8	Views as less innovative; requires technical skills.	Survey	Career Prestige Perception
Belief of ST negatively impacting workplace relationships.	39	The majority feel ST does not affect workplace dynamics.	Survey	Collectivist Cultural Value
ST impact workplace dynamics due to the nature of the role.	12	Stress and role-specific challenges.	Survey	Collectivist Cultural Value
Bayanihan affects career choices as teamwork is important.	25	Teamwork is important in choosing a workplace.	Survey	Collectivist Cultural Value
Preference for coding.	13	Simply stated "coding" or expressed passion for coding.	Survey	Career Prestige Perception
ST is seen as fun, interesting, has lighter workload, and easier than development.	-	Perceived pros of ST.	FGD	Career Prestige Perception
ST is seen as inferior to coding, repetitive, and has high learning curve.	-	Perceived cons of ST.	FGD	Career Prestige Perception
Pakikisama helps resolve issues easily between the tester and the developer.	-	Unique cultural value integrated in ST and SD dynamics.	FGD	Collectivist Cultural Value
Collaboration and teamwork are rooted in *bayanihan.*	-	Importance of local value in a workplace.	FGD	Collectivist Cultural Value
Utang na loob is there, but not a priority; personal interest is important.	-	Shift in traditional mindset.	FGD	Collectivist Cultural Value
Personal interest matters more than friends' opinions.	-	Personal interest as top motivator.	FGD	Collectivist Cultural Value
Work-life balance and environment matters.	-	Workplace dynamics is an important factor.	FGD	Collectivist Cultural Value

TABLE XII. EMERGENT ECONOMIC THEMES

Key Theme	Number of highlights	Notes	Data source	Framework Component
There is slow career progression in the role.	14	Perceived disadvantage.	Survey	Job Stability Expectations
ST is seen to provide less salary.	2	Concern on taking ST as a career.	Survey	Financial Necessity
ST provides limited career growth adding up to hesitation.	5	Concerns about long-term career growth and advancement prospects.	Survey	Job Stability Expectations
Salary, job-demand, and opportunities for growth are top priorities.	-	Top factors in choosing a career.	FGD	Financial Necessity
Gaining experience and building a strong portfolio are valued over simply following others.	-	Factor in choosing a career.	FGD	Job Stability Expectations
Job demand and economic needs are key factors.	-	When opportunities for this factor are present, a career is considered.	FGD	Job Stability Expectations
Opportunities for overseas work are an option.	-	Overseas work is not the priority, but when chances arise, it will be grabbed.	FGD	Financial Necessity
Careers with the promise of high salary are considered.	-	With reservations of the fact that fresh grad students earn minimum.	FGD	Financial Necessity

TABLE XIII. CONVERGENCE AND DIVERGENCE RESULTS MAP

Topic/Theme	Subtopic	Survey Findings	FGD Findings	Convergence/Divergence /Expansion
Interest Level	Likelihood to pursue	Median: likely to pursue.	Likely to pursue but with reservations.	Convergence
	IT Careers ranking	ST rank among other IT jobs: 3rd out of 5.	Software development is prioritized.	Divergence
Career Motivation	Job opportunities	Salary expectations: 0.323*	Salary is most prioritized.	Convergence
	Overseas work	0.163*	Overseas work is greatly considered.	Convergence
	Job market demand	0.639*	Whatever job/career is available, they will take.	Expansion
	Personal interest	0.883*	Driven by personal wants and needs.	Convergence
Curriculum Exposure	University preparedness	Unsure of how university prepares students.	Feels students are not well equipped for the profession.	Expansion
	Frequency of discussion	Much less frequent than any other IT jobs.	Agree, only shallow introductions.	Convergence
	Curriculum inclusion	Limited exposure.	Agree, not given highlight in course.	Convergence
	Testing education	0.057*	Less exposed = less interest.	Convergence/Expansion
Testing perceptions	Perceived pros	Easier than development	Testers only "test" that's why it's easier.	Convergence
		Necessary in development.	Agrees that it is necessary in development.	Convergence
	Perceived cons	Tedious work: Agree.	Agrees it is repetitive, manual work.	Convergence
		Requires expertise: Agree.	Agree, needs to keep up with learning curve.	Convergence
		Critics of developers: Agree.	Agree, it is the nature of the role.	Convergence
		Limited learning opportunities: Agree.	Disagree, there is always something to learn in the field.	Divergence
		Limited recognition: ranks #4.	Agree, developers are seen as superior.	Convergence
		Prone to workplace conflict ranks #7.	Agree, due to nature of the role.	Convergence
Social and cultural influences	*Utang na loob*	Median: neutral effect on ST choices.	Agree, leaning towards fulfillment of self.	Convergence
	Bayanihan	No impact on career choice but greatly impacts job satisfaction if present in workplace.	Would choose a career that values teamwork.	Expansion
	Pakikisama	No impact on career choices but it is very important to maintain good relationships with colleagues.	Wants a job that maintains workplace harmony.	Expansion
	Familial obligations	*Utang na loob*: Mean of influence: Unimportant.	Still thinks of giving back but not does not feel the pressure to do so.	Convergence
	Peer influence	0.107*	Would consider but to an extent.	Convergence/Expansion

*Note. *p-level of interest level to likelihood in pursuing a career in ST.*

However, students are open to pursuing careers in software testing when provided with proper education and exposure, challenging the assumptions about fixed preferences and suggesting that curriculum reform could greatly influence their career choices. An emerging theme among approximately 70% of participants indicated that economic factors, particularly salary expectations and job availability, have become primary determinants in their career selection, especially considering the current economic landscape in the Philippines. This pragmatic approach implies that if software testing careers align with financial stability and market demand, they become attractive options regardless of traditional cultural considerations. Focus group discussions consistently highlighted this economic prioritization, signaling a shift toward practical career decision-making over cultural or solely academic influences.

The integration of quantitative and qualitative findings reveals convergent patterns confirming salary expectations and overseas work considerations as primary career motivators, while diverging on the perceived value of learning opportunities in software testing. Table XIII identifies that cultural evolution among Filipino families demonstrates a shift from traditional authoritative practices toward supportive approaches that encourage individual autonomy, fundamentally transforming contemporary career decision-making processes. These insights establish that software testing's limited appeal stems from systemic issues, including inadequate curriculum exposure, economic pressures, and prestige perceptions, rather than inherent cultural resistance, suggesting these barriers could be addressed through coordinated educational and industry interventions. The interconnected nature of cultural, economic, and curricular factors creates a complex decision-making environment where students must balance personal interests, family obligations, financial necessities, and available opportunities when evaluating software testing career paths.

V. Discussion

Filipino IT students align with global trends favoring software development over testing, but show distinct cultural adaptations. Traditional values like *utang na loob* have evolved from primary to secondary motivators, with students asserting personal autonomy: "For me, my parents never forced me to choose what they wanted." Unlike international studies showing strong parental influence, Filipino students demonstrate a generational shift toward individualistic decision-making, prioritizing economic stability over family expectations: "…it's not my priority when it comes to choosing a career, given the fact their satisfaction... wouldn't feed me or wouldn't pay my bills".

Pakikisama emerges as crucial in workplace dynamics rather than career selection, particularly in software testing environments requiring diplomatic collaboration. While global studies emphasize prestige concerns, Filipino students additionally highlight economic necessities and limited curriculum exposure as decisive factors, reflecting the Philippines' developing economy, where financial stability takes precedence over career prestige. Cultural factors serve more as workplace compatibility considerations than career determinants, suggesting economic realities and educational limitations supersede traditional cultural expectations in contemporary Filipino career decision-making.

The study shows that Filipino IT curricula provide limited and superficial exposure to software testing, with participants reporting that ST is only briefly mentioned without comprehensive discussions or hands-on experience. This contrasts with more developed educational systems, where software testing is integrated throughout the software development lifecycle curriculum with practical projects and industry partnerships. The lack of structured ST education in Philippine institutions creates knowledge gaps that prevent students from recognizing software testing as a viable and attractive career path.

The primary barriers preventing Filipino IT students from pursuing software testing careers include insufficient academic exposure, limited curricular focus, and a lack of comprehensive knowledge about the field's opportunities and requirements. Economic considerations such as salary expectations and job availability take precedence over career exploration, particularly given the current Philippine economic conditions. Educational institutions fail to provide adequate career guidance and real-world exposure to software testing, leaving students with misconceptions about the field's potential and professional growth opportunities.

Filipino cultural values such as *pakikisama*, *bayanihan*, and *utang na loob* demonstrate nuanced influences on software testing career perceptions, with participants acknowledging their importance in workplace dynamics but not as primary career decision factors. While *pakikisama* facilitates harmonious tester-developer relationships and *bayanihan* promotes collaborative problem-solving in testing environments, these cultural elements contrast with Western individualistic approaches that emphasize direct communication and personal achievement. The traditional concept of *utang na loob* as a career motivator has diminished among contemporary Filipino IT students, who prioritize personal goals and economic stability over family obligations, reflecting a shift from collectivist to more individualistic career decision-making patterns.

VI. Conclusion and Recommendations

This study reveals that Filipino IT students' disinterest in software testing careers stems from limited academic exposure, superficial curricular coverage, and prioritization of economic factors over career exploration. Cultural values like *pakikisama*, *bayanihan*, and *utang na loob* primarily influence workplace dynamics rather than career choice itself. The research demonstrates a generational shift, where contemporary students prioritize personal fulfillment and economic stability over traditional familial obligations, making career decisions based on salary potential, market demand, and personal interests. This study contributes to career development literature by employing a multidimensional, culturally sensitive framework that bridges quantitative and qualitative findings, extending existing models by revealing how cultural, economic, and academic contexts interact in collectivist settings.

Based on these findings, comprehensive reforms are necessary to address the identified barriers and enhance students' interest in software testing careers. The following

recommendations are organized into three critical areas that require coordinated implementation across multiple stakeholders to effectively transform how software testing is perceived, taught, and integrated into IT education.

A. Educational Reform

- Implement comprehensive curriculum reforms, integrating hands-on software testing modules throughout IT programs
- Mandate CHED integration of software testing courses beginning in the second-year IT programs nationwide
- Develop faculty development programs to enhance testing pedagogy and industry-aligned teaching methodologies

B. Industry Partnership and Policy Development

- Establish strategic industry partnerships for internships, mentorship programs, and co-designed curricula that bridge academic learning with real-world testing practices
- Develop collaborative research initiatives where academia and industry jointly address software quality challenges, providing students with exposure to cutting-edge testing methodologies
- Implement targeted career guidance programs highlighting testing opportunities, competitive compensation, and professional growth trajectories through industry practitioner engagement

C. Future Research Directions

- Conduct longitudinal studies examining evolving student perceptions as curriculum reforms and industry partnerships are implemented.
- Validate the revised CEC-IRTC model through empirical testing in different cultural and educational contexts.
- Investigate the effectiveness of specific industry-academia collaborative models in shifting student career preferences toward software testing.

The proposed reforms require coordinated implementation across multiple stakeholders, with industry-academia collaboration serving as the critical bridge between curriculum development and workforce readiness. CHED curriculum integration provides the foundational policy framework, while sustained industry partnerships ensure that educational reforms remain responsive to market demands and provide students with authentic pathways into software testing careers. This collaborative approach transforms software testing education from an isolated academic concern into a shared responsibility between educational institutions and the IT industry, ultimately strengthening the software quality ecosystem in the Philippines.

REFERENCES

[1] Akosah-Twumasi, P., Emeto, T. I., Lindsay, D., Tsey, K., & Malau-Aduli, B. S. (2018). A systematic review of factors that influence youths' career choices—the role of culture. Frontiers in Education, 3, Article 58. https://doi.org/10.3389/feduc.2018.00058

[2] Alampay, L. P. (2024). Cultural values, parenting and child adjustment in the Philippines. International Journal of Psychology. Advance online publication. https://doi.org/10.1002/ijop.13117

[3] Angelo, F. A. (2019). Skills mismatch hampers Philippine economic growth, experts say. PIDS. https://www.pids.gov.ph/details/news/in-the-news/skills-mismatch-hampers-philippine-economic-growth-experts--say

[4] Bakke, I. B. (2021). Career and cultural context: collective individualism, egalitarianism and work-centrality in the career thinking of Norwegian teenagers. British Journal of Guidance & Counselling, 51(2), 282-297. https://doi.org/10.1080/03069885.2021.1872767

[5] BPAP (Business Process Association of the Philippines). (2020). Philippine IT-BPO industry statistics. Manila, Philippines.

[6] Bustos, R. (2024). Bridging the gap: Aligning higher education priorities with the shifting job landscape in the Philippines. Recoletos Multidisciplinary Research Journal, 12(1), 9–24. https://doi.org/10.32871/rmrj2412.01.02

[7] Capretz, L. F., & Gilal, A. R. (2022). What Pakistani computer science and software engineering students think about software testing? 574–575. https://doi.org/10.1109/APSEC57350.2022.00087

[8] Capretz, L. F., Basri, S., Adili, M., & Amin, A. (2020). What Malaysian software students think about testing? In ICSE '20: Proceedings of the 42nd International Conference on Software Engineering (pp. 1-2). Association for Computing Machinery. https://doi.org/10.1145/3387940.3392167

[9] Capretz, L. F., Waychal, P. K., Jia, J., & Basri, S. (2021). Comparing the popularity of testing careers among Canadian, Indian, Chinese, and Malaysian students. In AISS '21: Proceedings of the 3rd International Conference on Advanced Information Science and System (pp. 1-6). Association for Computing Machinery. https://doi.org/10.1145/3503047.3503091

[10] Commission on Higher Education (CHED). (2018). CMO No. 25, Series of 2015: Policies, standards, and guidelines for bachelor of science in computer science (BSCS) program. Manila, Philippines.

[11] Creswell, J. W., & Plano Clark, V. L. (2011). Designing and conducting mixed methods research (2nd ed.). SAGE Publications.

[12] Deak, A., Tor Stålhane, & Cruzes, D. S. (2013). Factors Influencing the Choice of a Career in Software Testing among Norwegian Students. https://doi.org/10.2316/p.2013.796-032

[13] Hassan, M. U., Alaliyat, S., Sarwar, R., Nawaz, R., & Hameed, I. A. (2023). Leveraging deep learning and big data to enhance computing curriculum for industry-relevant skills: A Norwegian case study. Heliyon, 9(4), e15407. https://doi.org/10.1016/j.heliyon.2023.e15407

[14] Lehtinen, T. O. A., Mäntylä, M. V., Vanhanen, J., Itkonen, J., & Lassenius, C. (2014). Perceived causes of software project failures – An analysis of their relationships. Information and Software Technology, 56(6), 623–643. https://doi.org/10.1016/j.infsof.2014.01.015

[15] Mowad, A. M., Fawareh, H., & Hassan, M. A. (2022). Effect of using continuous integration (CI) and continuous delivery (CD) deployment in DevOps to reduce the gap between developer and operation. In 2022 International Arab Conference on Information Technology (ACIT) (pp. 1–8). IEEE. https://doi.org/10.1109/acit57182.2022.9994139

[16] Nagtalon-Ramos, J. (2019). "Pakikipagkapwa and utang na loob": The influence of family and culture in graduate degree attainment among Filipino American nurses. Journal of Nursing Practice Applications & Reviews of Research, 10(1). https://doi.org/10.13178/jnparr.2020.10.01.1007

[17] Pascual, N. (2014). Factors affecting high school students' career preference: A basis for career planning program. International Journal of Sciences: Basic and Applied Research, 16(1), 1–14.

[18] Philippine Statistics Authority. (2022). 2022 Annual Survey of Philippine Business and Industry: Information and Communication Final Results. Manila, Philippines.

[19] Rafanan, R. J., De Guzman, C. Y., & Rogayan, D. V., Jr. (2020). Pursuing STEM careers: Perspectives of senior high school students. Participatory Educational Research, 7(3), 38–58. https://doi.org/10.17275/per.20.34.7.3

[20] Souza, R. E. C., Santos, R. S., Capretz, L. F., Sousa, M., & Magalhaes, C. V. (2022). Roadblocks to attracting students to software testing careers: Comparisons of replicated studies. In Communications in Computer and Information Science (pp. 127–139). Springer. https://doi.org/10.1007/978-3-031-14179-9_9

[21] Thomas, P., John, D., & Thomas, L. (2024). Factors influencing career decision of undergraduate and postgraduate students: An Indian context. Higher Education, Skills and Work-Based Learning, 14(1), 95–108. https://doi.org/10.1108/HESWBL-09-2022-0189

[22] Vautero, J., Silva, A. D., & Taveira, M. C. (2021). Family influence on undergraduates' career choice implementation. International Journal for Educational and Vocational Guidance, 21(3), 451–470. https://doi.org/10.1007/s10775-020-09453-0

[23] Yasin, A. (2020). What is the role of culture and family influence on career decisions of Asian tertiary students? International Journal of Academic Research in Business and Social Sciences, 10(3), 224–238. https://doi.org/10.6007/IJARBSS/v10-i3/7036

Pedestrian Navigation System with 3D Map and Charging Server Based on Steganography

Kohei Arai
Graduate School of Science and Engineering, Saga University, Saga City, Japan

***Abstract*—A pedestrian navigation system with a steganography-based 3D map and billing server is proposed. The proposed system includes a server system that provides topographical maps and navigation information to both pedestrians and vehicles. When using the proposed system, necessary images of cross sections, intersections, or points of interest can be automatically obtained, similar to the Street View feature in Google Maps. Users can post photos taken with their camera phones and earn points if the photos are marked as posted. While the proposed system incurs a usage fee, these points can be used to reduce the subscription fee. If the quality of an image is superior to that of a previously archived image, the new image overwrites the previous one. The billing system for the system usage fee incorporates digital steganography for security reasons. This prevents the leakage of user information and other data. Through experiments, it is confirmed that the proposed system works well.**

Keywords*—Pedestrian navigation; steganography; client-server system; geographic information; system; GIS

I. Introduction

The proposed system relates to a map information providing system that provides map information, particularly to a map information providing system that targets map information that is three-dimensional information [1]. As a background art of the proposed system, there is a position information management system disclosed in Japanese Patent Laid-Open No. 2003-111128. The position information management system of this background art is a data storage unit for storing map data and data for receiving landscape image data of a landscape around a user's information communication terminal from the information communication terminal via a communication line.

A configuration including a receiving unit, a current position specifying unit that specifies a current position where the information communication terminal is located based on the landscape image data received by the data receiving unit, and the map data stored in the data storage unit. According to the position information management system of the background art, when the position information from the GPS satellites is used without using the position information from the GPS satellites or the position information by the three-point survey method from a plurality of base stations in the vicinity.

The current position of the information communication terminal can be specified with the same high position accuracy. Further, the following technical matters are also disclosed in the publication.

- Specify the position based on the orientation data, in addition to the landscape image data.
- Identify the location using the area information of the base station.
- Update stored map data based on landscape image data.
- Navigate to the destination specified by the user.

The position information management system of background art cannot perform position identification using GPS in an environment such as a building or an underground mall, where it is difficult to receive radio waves from GPS satellites. In view of the above, the present invention has been made for the purpose of specifying the current position of the information communication terminal with the same high position as in the case of using GPS. However, the position information management system of the background art is configured as described above and has the following problems.

1) The map data is updated using the landscape image data captured by the user's information communication terminal, but the landscape image data is mistakenly updated because GPS is not used.

2) Landscape image data taken by the user's information communication terminal is taken at the user's discretion, and the landscape image data as planned by the system side cannot be acquired, resulting in poor-quality map data overall.

3) The scenery image data is repeatedly updated every time when it is uploaded to the system side, the load on the system side increases, and the scenery does not change frequently, so there is a lot of waste of updating itself.

The proposed system has been made to solve the above problems, and an object thereof is to provide a map information system in which image data relating to map information is appropriately updated.

In the following section, related research works are described. The proposed system is described in the next section, followed by some experiments. The conclusion is described with some discussions.

II. Related Research Works

An introduction to map matching for personal navigation assistants is provided [2]. Also, measurement and processing of indoor GPS signals using one-shot software receiver is conducted and well reported [3].

On foot navigation: continuous step calibration using both complementary recursive prediction and adaptive Kalman filtering is proposed [4]. Meanwhile, advanced integration of Wi-Fi and inertial navigation systems for indoor mobile

positioning is proposed and experienced [5].

Performance analysis of MEMS based pedestrian navigation systems is made [6]. On the other hand, 3D map creation with vehicle-mounted single stereo camera, reducing the influence due to vehicle motions, is proposed and validated its performance [7].

A moving domestic robot control method based on creating and sharing maps with shortest path finding and obstacle avoidance is proposed and evaluated its performance [8]. Meanwhile, color radio-map interpolation for efficient fingerprint Wi-Fi-based indoor location estimation is proposed [9].

Mapping and simulating watershed performances based on GIS, remote sensing, and SWAT are conducted and well reported [10]. On the other hand, communication links and evacuation/navigation information services for the tsunami warning system are introduced and experienced [11].

A seamless location measurement system with WiFi beacon utilized and GPS-based systems in both Indoor and outdoor receiver location measurement is proposed and evaluated for its performance [12]. Real-time monitoring visual impaired person's geolocation and walking state activity using GPS is also proposed [13].

Monitoring of visually impaired person movement using WEMOS di mini-GPS is also proposed [14].

As for digital steganography technology, a noble method for data hiding using steganography, discrete wavelet transformation, and cryptography, Triple Data Encryption Standard: TDES is proposed.

III. Proposed Method

A. System Concept

A map information providing apparatus according to the proposed system is an apparatus that transmits three-dimensional map information in response to a request from a mobile communication terminal equipped with a GPS receiver, and a three-dimensional map based on positioning information received from the mobile communication terminal. Means for reading the corresponding three-dimensional map information from the information storage means, means for reading the corresponding satellite image from the satellite image storage means based on the positioning information or the read three-dimensional map information, and imaging based on the positioning information or the read three-dimensional map information It includes a unit for reading the captured image from the image storage unit and a unit for pasting the satellite image and the captured image on the three-dimensional map information.

As described above, in the proposed system, the three-dimensional map information is read based on the positioning information from the mobile communication terminal, and the satellite image and the captured image are read as the images required for the three-dimensional map information and the satellite image is included in the three-dimensional map information.

Since the captured image is pasted and provided to the user, the satellite image and the captured image are interpolated with each other to provide the three-dimensional map information having a highly accurate real image.

The mobile communication terminal is a terminal that is portable and capable of communication, and specifically corresponds to a mobile phone, a PHS, a PDA including a communication device, and a notebook computer including a communication device. The captured image is an image captured by a mobile phone with a camera, a digital camera, a vehicle-mounted camera, or the like.

The satellite image may be attached to the top surface of the building and the road, and the captured image may be attached to the side surface of the object having the side surface (the road normally does not have the side surface). Alternatively, the captured image may be pasted where there is a captured image, and the satellite image may be pasted to other places. It is also possible to adopt a configuration in which one of the images is pasted first and then the other image is pasted on top of it.

A map information providing system according to the proposed system is a map information providing system including a mobile communication terminal equipped with an imaging device and a GPS receiver, and a map information providing apparatus that transmits three-dimensional map information in response to a request from the mobile communication terminal. Then, the map information providing device reads the corresponding three-dimensional map information from the three-dimensional map information storage means based on the positioning information received from the mobile communication terminal, and the captured image storage means based on the positioning information or the read three-dimensional map information.

The mobile communication terminal includes means for reading the captured image from the device, means for pasting the captured image on the three-dimensional map information, and means for recording the captured image from the mobile communication terminal in the captured image storage means based on the positioning information. It includes means for displaying the outline of the object in the three-dimensional map information based on the three-dimensional map information during the imaging mode.

In this way, in the proposed system, when capturing an image with the mobile communication terminal, the outlines of objects such as buildings, traffic lights, and intersections in the three-dimensional map information are displayed based on the three-dimensional map information, and the user matches the outline. It is possible to perform imaging, the map information providing apparatus can store an appropriate captured image assumed by the map information providing side, and it is possible to provide the user with three-dimensional map information to which a high-quality captured image is pasted.

The outline does not have to display the outline of all objects. For example, only the outline of the object located in the center may be used. The outline corresponds to a part of the frame (preferably hidden line processing) when the three-dimensional model is displayed in the wire frame.

In the map information providing system according to the proposed system, if necessary, the map information providing device transmits the positioning to the mobile communication terminal when a new captured image is needed from the captured image situation stored in the captured image storage means. It further includes means for requesting image pickup based on information.

As described above, in the proposed system, when the map information providing device determines that a new captured image is necessary according to the captured image situation of the captured image storage means, the image is captured based on the positioning information transmitted from the mobile communication terminal. When it is determined that the user is at a required position, the mobile communication terminal of the user is requested to capture an image, so that the user can capture an image required by the map providing device while suppressing the cost, and it is possible to collect high-quality captured images by the outline.

Here, "from the captured image status stored in the captured image storage means" means that, for example, when there is no captured image at a certain location, when there is a captured image, but the capture date is old, or there is a captured image, but at a specific time. This corresponds to the case where the captured image is insufficient. It is also a matter that the device operator can set appropriately.

A map information providing apparatus according to the proposed system is a map information providing apparatus that transmits three-dimensional map information in response to a request from a mobile communication terminal equipped with an imaging device and a GPS receiver, as necessary, and received from the mobile communication terminal. A means for reading the corresponding three-dimensional map information from the three-dimensional map information storage means based on the positioning information, a means for reading the corresponding captured image from the captured image storage means based on the positioning information or the read three-dimensional map information, and three-dimensional map information.

A means for pasting the captured image, a means for performing a route search from the current position and the destination when the destination is set from the mobile communication terminal, positioning information from the mobile communication terminal, and a mobile phone at the time of capturing the captured image. It includes means for identifying the image pickup direction of the communication terminal and creating information for guiding in the route direction using the image pickup direction and the route direction.

In this way, in the proposed system, the imaging direction of the mobile phone at the time of imaging is specified from the positioning information from the mobile communication terminal and the captured image, and the route to the destination and the information for guiding from the imaging direction to the route direction are created, By receiving the guidance by the mobile phone and outputting the guidance, the user can be guided in the correct direction without having to compare the surrounding situation with the 3D map information. The information to be guided corresponds to the information of the arrow in the route direction, the image of the correct route direction, and the image from the imaging direction to the route direction (may be a discontinuous image or a panoramic image).

A map information providing apparatus, according to the proposed system, is a map information providing apparatus that transmits three-dimensional map information in response to a request from a mobile communication terminal equipped with an imaging device and a GPS receiver as necessary and received from the mobile communication terminal. A means for reading the corresponding three-dimensional map information from the three-dimensional map information storage means based on the positioning information, a means for reading the corresponding captured image from the captured image storage means based on the positioning information or the read three-dimensional map information, and three-dimensional map information.

A means for pasting the captured image, a means for the map information providing device to store the positioning information and the embedded information from the mobile communication terminal in the embedded information storage means, and a transmission to the mobile communication terminal at the request of the mobile communication terminal 3 It includes means for including embedded information in the area of the three-dimensional map information in the three-dimensional map information.

As described above, in the proposed system, the positioning information and the embedded information from the mobile communication terminal are stored in the embedded information storage means and related to the three-dimensional map information to be transmitted, as well as the three-dimensional map information in response to the request of the mobile communication terminal. By including the embedded information to be provided to the user, the information registered by another user can be shared by the user who receives the information. It is possible to construct a so-called peer-to-peer usage mode in which information is directly exchanged between an unspecified number of individuals.

The map information providing apparatus or the map information providing system according to the proposed system embeds the management information in a hidden format and the management information in the image to be transmitted to the mobile communication terminal by the map information providing apparatus, if necessary. And a means for recording the recorded image. As described above, in the proposed system, since the management information is embedded in the image transmitted to the mobile communication terminal in a hidden format, it is possible to discriminate against an illegally used image and to record the image in which the management information is embedded. Since it is present, it can be used as evidence that it is the information already provided.

The map information providing system according to the proposed system additionally includes, if necessary, means for subtracting the charging information of the owner of the mobile communication terminal when the captured image to be recorded in the captured image storage means is received from the mobile communication terminal. It is a waste.

As described above, in the proposed system, when the captured image storage means is constructed using the captured image from the mobile communication terminal, the billing information regarding the use of the map information providing device for the user who provided the captured image by subtracting, it is possible to take an incentive for the user to take an image using the mobile communication terminal. The above outline of the proposed system does not enumerate the features that are essential to the proposed system, and a sub-combination of a plurality of these features can also be the proposed system.

B. System Configuration

Fig. 1 is a system configuration diagram of a map information system according to this embodiment.

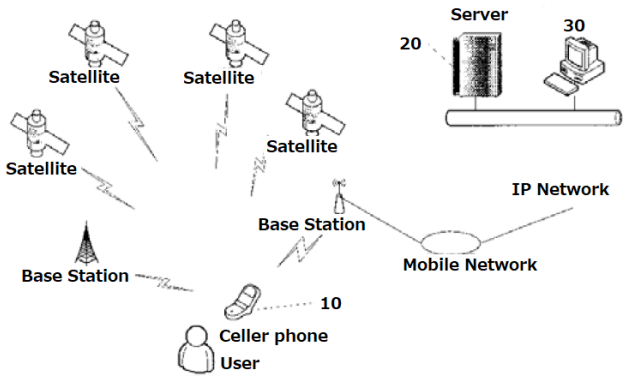


Fig. 1. System configuration diagram.

The map information system according to the present embodiment includes a mobile phone 10 with a GPS function and a camera function for receiving a radio wave from a GPS satellite and a reference station to identify a current position, a mobile phone 10, a base station, a mobile communication network, and an IP network. It is configured to include a server 20 that can communicate with the mobile phone 10 by transmitting the recorded three-dimensional map information to the mobile phone 10. In addition, a management computer 30 that operates and manages the map information system is also arranged.

Fig. 2 is a hardware configuration diagram of the components of the map information system according to this embodiment. Server 20 is built on a computer and has a hardware configuration such as a CPU (Central Processing Unit) 21, a main memory such as a DRAM (Dynamic Random Access Memory) 22, an external storage device HD (hard disk) 23, and a display device. A display 24, a keyboard 25, and a mouse 26, which are input devices, a LAN card 27, which is an expansion card for connecting to a network, a CD-ROM drive 28, and the like.

For example, the server program stored in the CD-ROM is duplicated (installed) on the HD 23, the server program is read into the main memory 22 as necessary, and the CPU 21 executes the program so that the server (device) Make up. Also, the management computer 30 has substantially the same hardware configuration as the computer on which the server 20 is constructed. Strictly speaking, the computer on which server 20 is constructed uses hardware with high performance and high reliability.

Most of the hardware configurations of the mobile phone 10 is the same as that of a computer, and instead of having no keyboard and mouse as input devices, a plurality of push buttons is provided on the surface of the main body, and a device and a camera function that enable mobile communication are realized. An example is a configuration in which a device that enables the realization and a device that receives a GPS radio wave and specifies the current position are mounted and are different.

The hardware configuration described here is an example, and it is obvious that those skilled in the art can make changes. Also, as for GPS, a plurality of methods such as single positioning, relative positioning, DGPS positioning, RTK-GPS positioning, etc., are prepared.

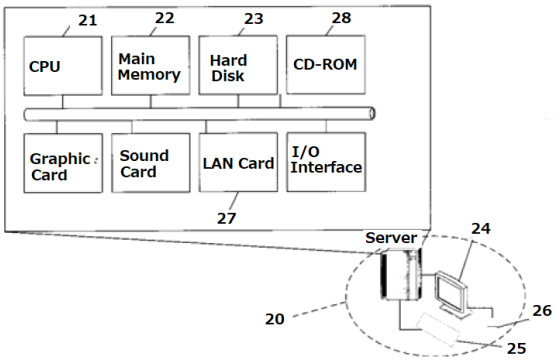


Fig. 2. Required functionality of the server.

Fig. 3 is a system configuration block diagram of the map information system according to the present embodiment. Server 20 is configured to have the following components.

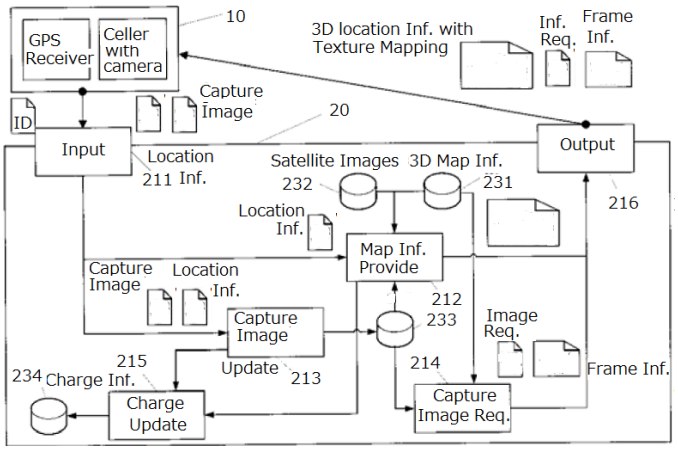


Fig. 3. System configuration.

Input means for capturing positioning information, captured images, etc. from the mobile phone 10. The corresponding three-dimensional map information is read from the three-dimensional map information storage unit 231 based on the captured positioning information, the satellite image is read from the satellite image storage unit 232 based on the positioning information, and the captured image storage unit 233 is read based on the positioning information.

A map information providing unit 212 that reads out the corresponding captured image and texture maps the satellite

image and the captured image on the three-dimensional map information. The captured image updating unit 213 that records the captured image in the captured image storage unit 233 based on the positioning information.

Picked image requesting means 214 for making an image pickup request to the user when updating is necessary from the picked-up image storage means 233 based on the positioning information.

Point updating means 215 for updating the charging information based on the information provided by the map information providing means to the user.

An output unit 216 that sends out three-dimensional map information, an image pickup request, frame information, and the like to the mobile phone 10. Next, the operation of the map information providing system according to this embodiment will be described. The operation according to the present embodiment is a three-dimensional map display operation of displaying three-dimensional map information on which a real image is mapped on the user's mobile phone 10, and a captured image captured by the user's mobile phone 10 as three-dimensional map information. A captured image update operation of storing a mapable image, and a captured image request operation of requesting a captured image user to map to three-dimensional map information, and storing a captured image captured in response to the request in a mapable manner with three-dimensional map information.

C. 3D Map Display Operation

Fig. 4 is a flowchart of the three-dimensional map display operation according to this embodiment. First, the user uses the push button of the mobile phone 10 to select the map information providing service. Then, the mobile phone 10 receives radio waves from the GPS satellite and the reference station and obtains positioning information (step 101). The obtained positioning information is sent to server 20 via the communication path established between the mobile phone 10 and the server 20. Server 20 (map information providing means 212) reads out the three-dimensional map information required for display from the three-dimensional map information storage means 231 based on this positioning information (step 111).

Server 20 (the map information providing unit 212) reads out the satellite image and the captured image required for mapping from the satellite image storage unit 232 and the captured image storage unit 233 based on the positioning information (step 112). Although the image is read out based on the positioning information, it goes without saying that the image may be read out based on the read three-dimensional map information.

The server 20 (map information providing means 212) maps the read satellite image and captured image to three-dimensional map information. Here, mapping can be performed using various mapping techniques such as texture mapping. As a rule, a satellite image is pasted on the upper surface of the 3D map model and a captured image is pasted on the side surface of the 3D map model.

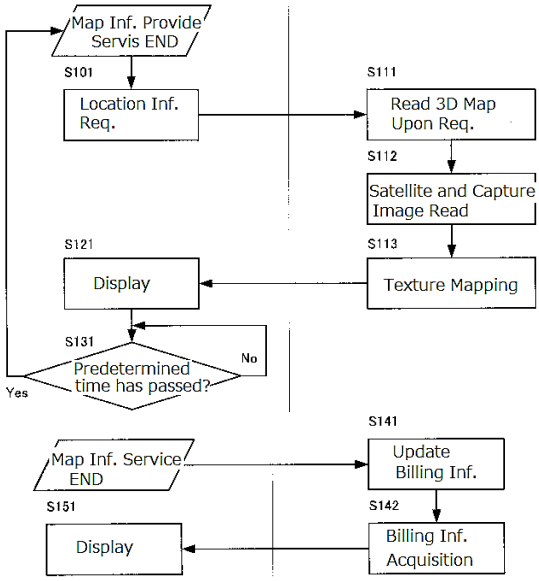


Fig. 4. Flowchart of the three-dimensional map display operation.

Since satellite images can also be obtained from the perspective direction, they may be attached to the side surface in some cases. The mapping here can, of course, be automated, but an operator can paste it instead. Server 20 sends the mobile phone 10 three-dimensional map information, in which the satellite image and the captured image are mapped. The mobile phone 10 fetches the three-dimensional map information and displays the three-dimensional model on the display.

The user can see the three-dimensional map information of the surrounding area. Here, the transmitted 3D map information may be information in a very limited range of the current position of the user, or may be 3D map information of a certain predetermined rectangular area. It is necessary to capture the 3D map information described above at any time according to the movement of the user. In the latter case, if the user moves to some extent, it is not necessary to capture new 3D map information within the rectangular area range.

The mobile phone 10 determines whether a predetermined time has passed from the previous step 101 (step 131), returns to step 101 if it determines that it has passed, and returns to step 131 if it determines that it has not passed. Return. Here, the three-dimensional map information to be fetched is updated every predetermined time, but the configuration may return to step 101 in accordance with the update of the positioning information. At this time, the positioning information may simply be updated, or when the moving distance from the previous step 101 exceeds a predetermined threshold value, the process may return to step 101.

Although the map information providing service is implemented in this way, when the user presses the push button of the mobile phone 10 and selects the end of the map information providing service, the service end signal is received and the server 20 (billing the updating means 215) adds the charging information (step 141). The billing information may be added based on the amount of the map providing service provided. The accounting information added by the server 20

(account updating means 215) is acquired (step 142). The acquired billing information is sent to mobile phone 10.

The mobile phone 10 displays the billing information on the display (step 151). The user can see the billing information. It should be noted that between the mobile phone 10 and the server 20, information (for example, a mobile phone number or a member ID assigned when registering to receive a service) that identifies the user of the mobile phone 10 is used as necessary. There is a need. It is necessary at least for billing renewal. The same applies below.

D. Captured Image Update Operation

Fig. 5 is a flowchart of the captured image update operation according to this embodiment.

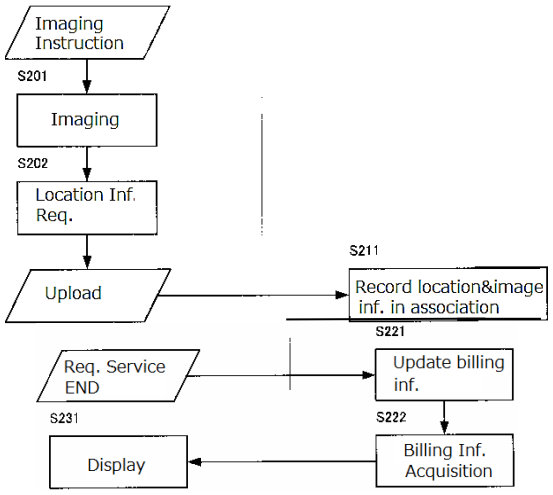


Fig. 5. Flowchart of the proposed imaging process.

First, the user uses the push button of the mobile phone 10 to select an image capturing instruction during the map information providing service (it may not be in service). The mobile phone 10 carries out imaging in accordance with the imaging instruction (step 201).

The mobile phone 10 obtains positioning information at the time of image pickup (step 202). The user uses the push button of the mobile phone 10 to select the upload of the captured image. The captured image and the positioning information are sent to Server 20. The server 20 (captured image updating means 213) stores the positioning information and the captured image in the captured image storage means 233 in association with each other (step 211). If a picked-up image of the same positioning information already exists, the picked-up image just stored is updated.

However, it is possible to retain the previously captured image, and it is possible to provide the three-dimensional map information in which the captured image of high quality is mapped by the confirmation of the operator. Instead of operator confirmation, noise determination processing for noise inclusion, scale determination processing to determine the scale of the captured image by edge processing of buildings, etc. From the image, the cutout processing to cut out the necessary image from the scale is automated. It can also be configured to.

When the user presses the push button of the mobile phone 10 and selects the end of the map information providing service, the service end signal is received, and in addition to the addition of the charge for the map information providing service, the server 20 (charge updating means 215) subtracts the billing information (step 221).

The billing information may be subtracted based on the number of uploaded captured images. The accounting information updated by the server 20 (account updating means 215) is acquired (step 222). The acquired billing information is sent to mobile phone 10. The mobile phone 10 displays the billing information on the display (step 231). The user can see the billing information. According to this captured image update operation, the captured image can be updated in a self-proliferating manner, and the work for maintaining the captured image can be significantly reduced.

E. Best Mode for Implementing the Proposed System

The proposed system can be implemented in many different forms. Therefore, it should not be interpreted only by the description of the embodiments below. In the embodiments, the system will be mainly described, but as will be apparent to those skilled in the art, the proposed system can be implemented as a program and method usable by a computer. Also, the proposed system can be implemented in hardware, software, or software and hardware embodiments.

The program can be recorded in any computer-readable medium such as a hard disk, a CD-ROM, a DVD-ROM, an optical storage device, or a magnetic storage device. Further, the program can be recorded on another computer via the network. Fig. 6 is a flowchart of the captured image request operation according to this embodiment.

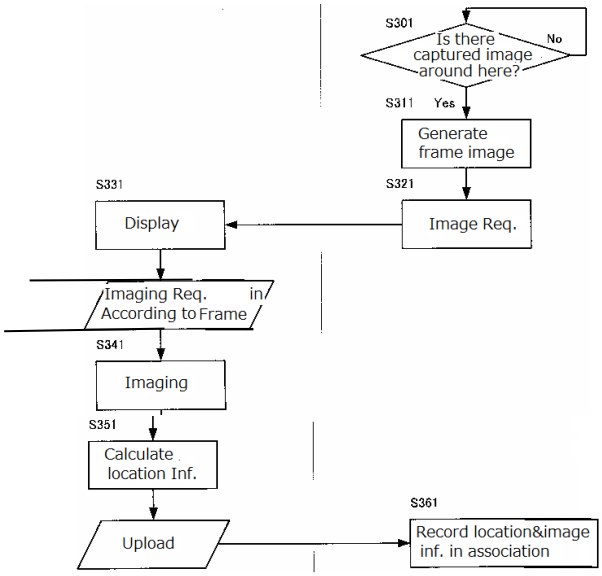


Fig. 6. Flowchart of the captured image request operation.

Even if at least one user can perform the captured image request operation when the 3D map information service is provided. Here, when no one is using the captured image requesting operation, the user who has registered by e-mail or the like, and the user who has used it can obtain the captured

image request operation. Specifically, when it is acceptable to be the target user of the captured image requesting operation, a message requesting access to this URL is transmitted.

Server 20 (captured image requesting means 214) determines whether an image around the current position is stored in the captured image storage means 233 using the positioning information of the user who is in service (step 301). If it is stored, the process returns to step 301. If it is not stored, server 20 (captured image requesting means 214) generates a three-dimensional model that clearly shows the required frame of the building or the like from the three-dimensional map information around the current position (this is the wire frame itself. Or it may be one that has been subjected to hidden line treatment, or one that has been rented with a white outer surface. In addition, a convenience store, a store such as a gas station, or an intersection that will serve as a mark is clearly indicated. It may be configured), and the direction in which imaging is required is specified.

The server 20 sends the generated three-dimensional model and the positioning information of the imaging position to the mobile phone 10 as an imaging request.

The mobile phone 10 displays the three-dimensional model on the display (step 331), and the user gives an instruction to image in a designated direction in accordance with the frame. The mobile phone 10 performs imaging (step 341). The mobile phone 10 obtains positioning information at the time of imaging (step 351). The user uses the push button of the mobile phone 10 to select the upload of the captured image. The captured image and the positioning information are sent to server 20. The server 20 (captured image update means 213) updates the positioning information and the captured image in association with each other, as in step 211 (step 361). Steps 221 to 231 are also executed in this captured image requesting operation.

Since this captured image requesting operation is often carried out frequently during the 3D map information providing service, whether the captured image requesting operation may be performed for the user in the 3D map information providing service in advance. It is preferable to have a configuration that allows selection. According to this captured image, requesting operation, captured images required for the map information providing service can be efficiently collected.

When it is desired to collect mainly the picked-up images of a predetermined area in an urban area, it is possible to set to collect for such an area. Further, here, the trigger of the captured image request operation is separated only by the determination in step 301 of whether there is a captured image, but the target of the captured image request operation is even when a predetermined period has elapsed from the image capture date and time of the captured image.

It is also possible to always keep the latest image. Further, although the image capturing time is not specifically specified, the captured image is different depending on the image capturing time. It is also more preferable to carry out. Note that the server 20 side during the captured image request operation generated the three-dimensional model in which the frame is clearly defined, but the mobile phone 10 side during the captured image update operation generates the three-dimensional model in which the frame is clearly displayed using the three-dimensional map information. It can also be configured.

Although it is not mentioned about the processing here, which functions are to be implemented on the client side and which functions are to be implemented on the server side in a client-server system etc., are almost always changed within the scope of so-called design change by those skilled in the art to be able to.

F. Navigation Procedure

Fig. 7 is a flowchart of navigation operation according to this embodiment. The user presses the push button of the mobile phone 10 to select the navigation. The mobile phone prompts the user to specify the destination by using the same GUI as the in-vehicle navigation. The user sets the destination. The mobile phone 10 requests the positioning information (step 401). The mobile phone 10 sends the destination and the positioning information to the server 20. The server 20 (route searching means 217) searches for a route from the destination and the positioning information (step 411).

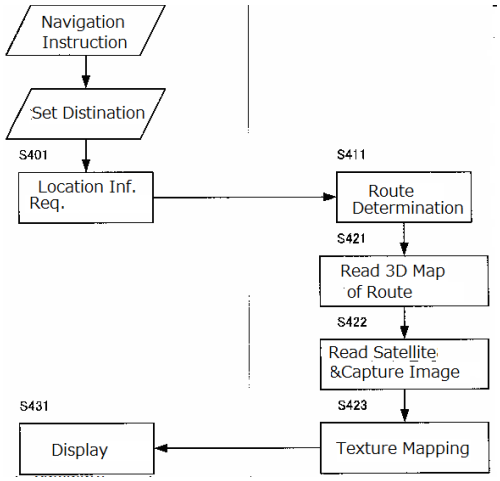


Fig. 7. Navigation procedure.

The route search is not an essential part of the proposed system, and a so-called person skilled in the art can apply a route search method that is already well known and will not be described in detail here. The server 20 (route searching means 217 and map information providing means 212) reads out the three-dimensional map information from the three-dimensional map information storage means 231 based on the route (step 421). The server 20 (the route searching means 217 and the map information providing means 212) reads the satellite image and the picked-up image from the satellite image storage means 232 and the picked-up image storage means 233 based on the route (step 422). The server 20 (route searching means 217 and map information providing means 212) attaches the satellite image and the captured image to the three-dimensional map information and further reveals the route (step 423). Revealing a route is generally a curve from the current position to the destination. The server 20 sends the three-dimensional map information to the mobile phone 10.

The mobile phone 10 displays three-dimensional map information on the display (step 431). The user can see a three-

dimensional map around the current position or can see a three-dimensional map other than the current position on the route. Here, the three-dimensional map information on the route is collectively sent from the current position from the server 20 and displayed on the mobile phone 10, but the three-dimensional map is updated based on the update of the positioning information accompanying the user's movement or a predetermined time. Information can be sent out at any time. In addition, the billing update can be performed in the same manner as in steps 141 to 151.

Fig. 8 is a flowchart of the direction confirmation operation according to this embodiment. The user instructs the mobile phone 10 to take an image during the navigation operation. The mobile phone 10 carries out imaging (step 501). The bay 10 requests the positioning information by carrying (step 511). The user presses the push button of the mobile phone 10 and selects the function of confirming the direction. The mobile phone 10 sends the captured image and the positioning information to the server 20.

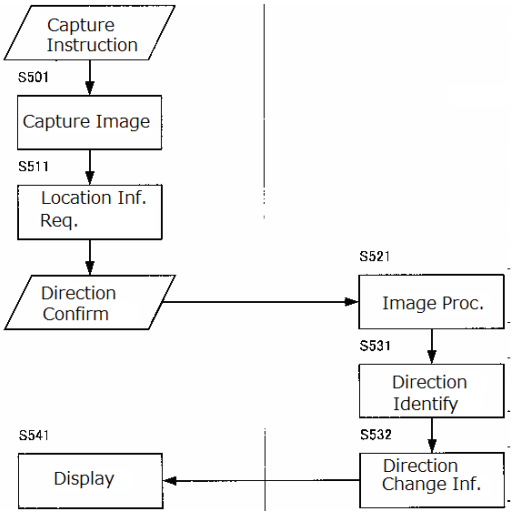


Fig. 8. Flowchart of the direction confirmation operation.

The server 20 performs image processing on the captured image to obtain an edge of a building or the like (step 521). Server 20 reads the three-dimensional map information from the three-dimensional map information storage unit 231 from the positioning information, compares the edges with the frame of the three-dimensional map information, and specifies the matching direction (step 531). The server 20 refers to the route of the three-dimensional map information for navigation sent to the user, identifies the correct direction on the route, and obtains direction change information from the direction identified in step 531 (step 532).

The direction change information includes a message for direction change, such as "How many times should the direction be changed clockwise from the current direction?" Several 20 sends the direction change information to the mobile phone 10. The mobile phone 10 displays direction change information on top of the currently displayed three-dimensional information for navigation (step 541). In addition, the billing update can be performed in the same manner as in steps 141 to 151.

According to this direction confirmation operation, the user visually compares the navigation with the image facing, as compared to the case where the user visually checks the navigation display and the image in the direction facing to reach the destination. No need. If the captured image is not mapped, and if no direction can be identified by the landscape of the same building, it points to an appropriate direction, which is highly convenient.

G. Navigation Procedure

The map information system according to the present embodiment is as follows: The map information system according to the present embodiment is configured in the same manner as the map information system according to the first embodiment, receives an image request from a certain user, and makes an image request to a user in an area satisfying the image request. This is different from the configuration in that the image requesting means for sending out the image and the steganography means for embedding the management identification information in the image are newly provided.

The steganography technique used in steganography is a well-known technique, and so-called persons skilled in the art can embed embedded information in a target image using various steganography techniques. Digital steganography technology can be used instead of steganography technology [15].

IV. Experiment

A. Direction Display

In each of the above-described embodiments, when the mobile phone 10 is displaying three-dimensional map information on the display, for example, when the direction key is pressed, the mobile phone 10 rotates in the assigned direction by the amount of the pressing. It can be configured to display. Specifically, when the front is displayed, by pressing the left direction key for a predetermined time, the landscape changes counterclockwise during the pressing time.

B. Three-Dimensional Album

In the map information system according to the second embodiment, during navigation, images and moving images captured by the user's mobile phone are sent to the server 20 at any time together with the positioning information, and the image for the navigational three-dimensional map information is displayed. Alternatively, the moving image may be embedded based on the positioning information and sent to the mobile phone 10 as a three-dimensional album.

C. Embedded Information for 3D Map Information

In the map information system according to the first embodiment, the user uploads the captured image in the captured image update operation and uses the uploaded captured image as an image to be attached to the three-dimensional model. Alternatively, for example, when a restaurant is specified, the captured image of the restaurant interior, the captured image of the food provided, and a description of the restaurant may be sent together with the positioning information, and the user can display the embedded information.

When requested, the three-dimensional map information including the embedded information may be displayed. Furthermore, if you create a site that displays only embedded information for each genre in a menu format and the user who accesses the site selects "3D map information", various services for 3D map information of this system will be provided. It can also be configured too.

V. Conclusion

A pedestrian navigation system with a 3D map and a charging server based on steganography is proposed. The server system, which provides topographic maps and navigation information for pedestrians as well as automobiles, is proposed. Required images of cross sections, intersections, or points of interest can be automatically acquired from users when they use the proposed system. Users get beneficial points when they submit photos acquired with their mobile phone camera, so that they intend to submit. If the quality of the image is better than previously archived image, then the new image is overwritten over the previous one. Steganography is featured for security reasons in the charging system for the use of the system.

Although the presently proposed system has been described by the above respective embodiments, the technical scope of the presently proposed system is not limited to the scope described in the embodiments, and various changes or improvements can be added to these respective embodiments. The embodiment with such changes or improvements is also included in the technical scope of the proposed system.

Future Research Works

Further experiments are required for validation of the proposed method and system in actual situations. Research is underway in automated collection and updating of real-time barrier information using smartphone sensors and AI technology. Real-time direction estimation using deep learning and image-based correction algorithms is also evolving. Infrastructure-free navigation technologies, such as high-precision autonomous positioning methods using shoe-mounted inertial sensors, are gaining attention.

Acknowledgment

The author would like to thank Prof. Dr. Hiroshi Okumura and Prof. Dr. Osamu Fukuda of Saga University for their valuable comments and suggestions.

References

[1] Inventor: Kohei Arai, Applicant: Saga University, PCT/JP2005/20044,(2005), 2005.

[2] D. Bernstein, A. Kornhauser, "An introduction to map matching for personal navigation assistants', Report of New Jersey TIDE Center, New Jersey Institute of Technology, 1996.

[3] G. Lopez-Risueño, G. Seco-Granados, "Measurement and Processing of Indoor GPS Signals Using One-Shot Software Receiver," in Proceedings of the IEEE Vehicular Technology Conference Spring 2005, 2005.

[4] Q. Ladetto, "On foot navigation: continuous step calibration using both complementary recursive prediction and adaptive Kalman filtering", Proc. ION GPS 2000.

[5] Frédéric Evennou and François Marx, "Advanced Integration of WiFi and Inertial Navigation Systems for Indoor Mobile Positioning", EURASIP Journal on Applied Signal Processing, Vol. 2006, Article ID 86706, 2006.

[6] D. Kubrak, C. Macabiau, M. Monnerat, "Performance Analysis of MEMS based Pedestrian Navigation Systems", Proceedings of ION GNSS 2005.

[7] Kohei Arai, 3D map creation with vehicle mounted single stereo camera reducing influence due to vehicle motions, International Journal of Research and Reviews on Computer Science, 2, 4, 1056-1062, 2011.

[8] Kohei Arai, Moving domestic robot control method based on creating and sharing maps with shortest path finding and obstacle avoidance, International Journal of Advanced Research in Artificial Intelligence, 2, 2, 23-28, 2013.

[9] Kohei Arai, Herman Tolle, Color radio-map interpolation for efficient fingerprint WiFi-based indoor location estimation, International Journal of Advanced Research in Artificial Intelligence, 2, 3, 10-15, 2013.

[10] Sandhyavitri Ari, Yusa Muhamad, Fajra Mardan, Sutikno Sigit, Kohei Arai, Mapping and Simulating Watershed Performances Based on GIS, Remote Sensing and SWAT, Proceedings of the International Conference on Advanced Civil and Environmental Engineering, ICAnCEE 2018, Bali, Indonesia., 2018.

[11] Kohei Arai, Communication links and evacuation/navigation information services for tsunami warning system, Proceedings of the 1st International Workshop on Knowledge Cluster Systems,2007.

[12] Kohei Arai, Seamless location measurement system with WiFi beacon utilized and GPS based systems in both of Indoor and outdoor receiver location measurement, International Journal of Advanced Research on Artificial Intelligence, 4, 7, 35-40, 2015.

[13] Anik Nur Hardayani, Dyah Lestari, Harits Rosyd, Mohammad Efendi, Aji Wibawa, Rosa Asumara, Kohei Arai, Real time monitoring visual impaired person geolocation and walking state activity using GPS, Proceedings of the SAIN 2019.

[14] Anik Nur Hardayani, Dyah Lestari, Harits Rosyd Ar, Rosa Asamura, Kohei Arai, Monitoring of visual impaired person movement using wemos di mini GPS, Proceedings of the 6th International Conference on Electrical, Electronics and Information Engineering, ICEEIE, 2019.

[15] Cahya Rahmed Kohei Arai, Arief Prasetyo, Noriza Arigki, Noble Method for Data Hiding using Steganography Discrete Wavelet Transformation and Cryptography Triple Data Encryption Standard: DES, IJACSA, 9, 11, 261-266, 2018.

Author's Profile

Kohei Arai. He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he was award committee member of ICSU/COSPAR. He wrote 190 books and published 750 journal papers as well as 550 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.html

Comparative Evaluation of CNN Architectures for Skin Cancer Classification

Taopik Hidayat, Nurul Khasanah, Elly Firasari, Laela Kurniawati, Eni Heni Hermaliani
Faculty of Information Technology, Universitas Nusa Mandiri, Jakarta, Indonesia

***Abstract*—Skin cancer is one of the fastest-growing health problems worldwide. Early and accurate diagnosis is essential for improving treatment success and patient survival. However, many previous studies have focused on single CNN architectures or limited datasets, resulting in models with restricted generalizability. To address this gap, this study presents a comparative evaluation of three deep learning architectures (DenseNet169, MobileNetV2, and VGG19) for automatic classification of benign and malignant skin cancers using dermoscopic digital images. A total of 10,000 images were compiled from three public Kaggle datasets, preprocessed through resizing and data augmentation, and trained using transfer learning based on ImageNet weights. Two data split schemes (60:20:20 and 80:10:10) were applied to assess model robustness. Experimental results show that DenseNet169 achieved the highest test accuracy of 90.7 per cent, while MobileNetV2 was the fastest with an inference time of 16 seconds. These findings highlight the tradeoff between accuracy and computational efficiency and support the use of deep learning models, particularly DenseNet169 and MobileNetV2, in the development of real-time AI-assisted skin cancer diagnostic systems.**

Keywords*—Artificial intelligence; convolutional neural network; deep learning; dermoscopic images; skin cancer classification

I. Introduction

Skin cancer is a common type of cancer that has been increasing rapidly year after year [1]. According to WHO data, each year, there are over two million cases of skin cancer diagnosed, the majority of which are benign, such as actinic keratosis and basal cell carcinoma [2]. Malignant skin cancers, such as melanoma, continue to pose a substantial danger due to their high metastatic potential and death rate [3]. Since the speed at which the lesion can be identified and treated has a significant impact on treatment outcomes, early diagnosis is an essential step in lowering melanoma mortality. Unfortunately, the standard skin cancer detection procedure continues to rely mainly on visual inspection by a dermatologist, which is subjective and requires extensive clinical experience, rendering it prone to errors and inconsistencies [4].

Automation attempts in medical diagnosis have become the subject of significant research, particularly when it comes to medical image processing [5], [6], [7], [8], [9], [10], [11]. The use of deep learning, which excels in extracting visual features and accurately classifying objects, is one of the popular methods [12]. Deep learning can learn visual patterns from picture data from start to finish, eliminating the need for human feature extraction [13]. Deep learning can distinguish tiny distinctions between skin lesion types that are often not visible to the human eye, making it a promising technology to support objective, rapid, and efficient clinical judgments [14], [15], [16].

Several studies in classifying skin lesions in dermatoscopic images have successfully demonstrated using CNN models. Previous studies [17] proposed a hybrid deep learning model to improve the accuracy of skin cancer classification by combining ConvNeXtV2 blocks and a separable self-attention mechanism, especially in differentiating between benign and malignant lesions that share a similar appearance. This model was shown to be more superior than the other ten CNN and ViT models with an accuracy of 93,48% and only 21,92 million parameters, and also efficient for clinical applications. Furthermore, deep learning methods with ResNet, VGG16, and AlexNet models were applied to the HAM10000 dataset for skin cancer classification, with improvements through data augmentation and imbalanced data addition, resulting in accuracies of 92,9%, 98,4%, and 88,3%, respectively, where the improved VGG16 model was selected as the best model to support early detection of skin cancer [18]. Further research [19] developed an automatic skin cancer detection system based on the DenseNet169 model trained using the Skin Cancer: Malignant versus Benign dataset, and successfully achieved a high accuracy of 89.7%, surpassing several conventional models and potentially supporting medical personnel in more accurate and efficient early diagnosis. Six deep learning models, including EfficientNet (B0, B1, B2) and MobileNet (V2, V3-Small, V3-Large), were used in the study [20]. Each model uses the ISIC dataset to classify skin cancer. The best results were obtained from MobileNet-V3-Large with an accuracy of 89.41%, a recall of 90.59%, and an F1-score of 89.53%, accelerating diagnosis and improving accuracy [20]. Another study focused on the use of the MobileNetV2 model for digital image-based skin cancer classification. After training and evaluation, the MobileNetV2 model successfully achieved an accuracy of 85% in distinguishing between cancerous and non-cancerous skin lesion images [21]. In another study [22], several pre-trained models are applied in the deep learning methods, one of which was VGG19. The VGG19 model achieved an accuracy of 87% using the HAM10000 dataset after the augmentation and fine-tuning process, showing quite good performance in classifying skin cancer. While the study [23] enhanced the pre-trained VGG19 model by adding max pooling and dense layers to improve the skin cancer prediction capability. The improved VGG19 model successfully achieves an accuracy of 88% in skin cancer classification.

Referring to the background, the purpose of this study is to evaluate and compare the performance of seven deep learning architectures in classifying skin cancer into two main categories,

namely benign and malignant. The models used are MobileNetV2, InceptionV3, Xception, DenseNet169, ResNet50, VGG16, and VGG19, each of which represents a variation in architectural complexity and computational efficiency. The dataset used was obtained from the Kaggle platform, which consists of 10,000 digital images that have been labeled according to the type of lesion. This study uses a transfer learning approach, where all models are initialized with weights from pre-training on ImageNet, then the medical data is fine-tuned to match its characteristics.

The main purpose of this study is to achieve an optimal deep learning model to classify skin cancer based on evaluation metrics, including recall, precision, accuracy, and F1-score. The results of this study are expected to contribute in the advancement of digital picture development-based clinical decision support systems, especially in the diagnosis of skin cancer. In addition, this study also aims to encourage the use of artificial intelligence in the health sector more widely by considering the aspects of reliability, speed, and efficiency of medical classification systems. It is also expected to provide a thorough grasp of the strengths and the limitations of deep learning architecture in the classification of benign and malignant skin cancer by conducting a comprehensive evaluation of the performance of each model. In addition, the results of this study can be a starting point for the development of an intelligent diagnostic system that is more accurate, easily accessible, and able to support early detection of skin cancer effectively.

The novelty of this study lies in the comprehensive comparison of CNN architectures using multiple public datasets and dual data split schemes to assess both accuracy and inference efficiency in real-world skin cancer classification.

The remainder of this study is organized as follows: Section II reviews related studies on skin cancer classification using deep learning. Section III explains the materials and methods, including dataset composition, preprocessing, and CNN architectures. Section IV presents the experimental results and analysis, while Section V concludes the study with key findings and future directions.

II. Related Works

The previous related studies using digital image processing approaches and deep learning methods for image classification are summarized in Table I. Various deep learning models such as DenseNet169, DenseNet121, MobileNetV2, VGG19, ResNet50, Xception, InceptionV3, and EfficientNetB0 have been applied in these studies. The accuracy results of these models vary, ranging from 85% to 89%.

Table I presents a comparison of the results of several previous studies using a deep learning approach to classify and detect skin cancer. Most studies use a digital image-based approach with popular models including DenseNet, MobileNet, VGG, ResNet, Xception, Inception, and EfficientNet. An accuracy of 89.7% is the highest result using the DenseNet169 model [19], followed by the E-VGG16 model [24] and DenseNet121 [25], which each recorded an accuracy of 89%. Another study [22], [23] showed competitive results using VGG19 and its modified version, with accuracies of 87% and 88%. These results indicate that deep learning models are suitable for skin cancer classification with high accuracy results.

In addition, the MobileNetV2 model was used in two different studies with accuracy results of 85.6% and 85% respectively, indicating the efficiency of this model in a limited computing environment, even with a slight decrease in accuracy [20], [21]. ResNet50 also showed consistency with 87% accuracy in two separate studies [10], [26]. Further research used a combination of features from the Xception and InceptionV3 models to detect Monkeypox skin disease and obtained an accuracy of 85.9%, indicating the potential of the combination architecture approach [27]. Meanwhile, research using EfficientNetB0 recorded an accuracy of 87% [28], showing the superiority of a lightweight yet effective model. Overall, the data in this table indicates that various deep learning models have demonstrated competitive performance and are worth considering for image-based skin disease classification systems. However, few studies have systematically examined how variations in dataset composition and data split schemes affect the performance and generalizability of CNN architectures for skin cancer classification.

TABLE I. Related Works Summary

Author	Task	Method	Acc (%)
R. Pathania et al. [19]	Skin cancer detection	DenseNet169	89.7
O. Sahin et al [20]	Skin cancer classification	MobileNetV2	85.6
D. Moturi et al [21]	Melanoma skin cancer detection	MobileNetV2	85
I. Ahmad et al [22]	Skin cancer detection	VGG19	87
I. Kandhro et al [23]	Skin cancer detection and classification	E-VGG19	88
D. Albashish et al [26]	Melanoma skin cancer classification	ResNet50	87
N. Pratama et al [27]	Monkeypox skin disease detection	Combining feature Xception and InceptionV3	85.9
N. Khasanah et al [10]	Melanoma skin cancer detection	ResNet50	87
S. Matiray et al [28]	Skin cancer detection	EfficientNetB0	87
V. Anand et al [25]	Skin disease classification	DenseNet121	89
S. Mushtaq et al [24]	Skin cancer classification	E-VGG16	89

III. Materials and Methods

In this section, the stages of the methodology applied in the study are described. This study focuses on testing and comparing seven deep learning models with a transfer learning approach, using public data from the Kaggle platform to classify benign and malignant skin cancers. The entire process is designed systematically, starting from the data input stage, pre-processing, model training, to evaluation of results, as shown in the workflow illustration in Fig. 1.

In this study, the primary data used consisted of 10,000 skin cancer images obtained from the Kaggle platform. The initial stages of the study included data pre-processing, which consisted of image resizing and augmentation applications to increase the diversity and quality of training data. Furthermore, this study tested and compared the performance of seven deep learning model architectures, namely, MobileNetV2, InceptionV3, DenseNet169, ResNet50, Xception, VGG16, and VGG19. Fig. 1 provides an overview of the main stages in the research process. In contrast, Fig. 2 shows the complete research stages of this research, which consist of collecting the data, processing the data itself, continuing to the training process with several training models, the evaluation process, and finally drawing conclusions. Fig. 3 provides a detailed visualization of the training and implementation stages of the model in this study, showing an example of the transfer learning application process flow on the seven models.

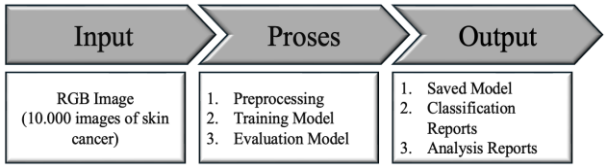


Fig. 1. Overview of the general stages of the research.

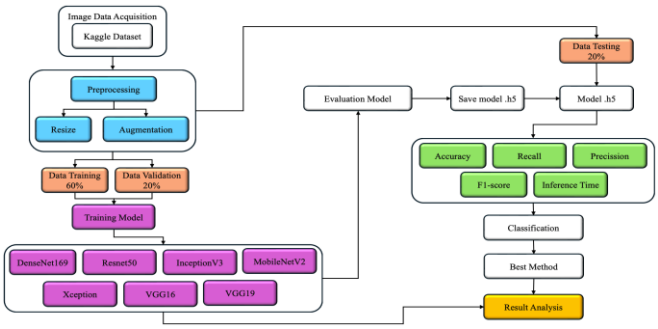


Fig. 2. Flowchart of the research stages.

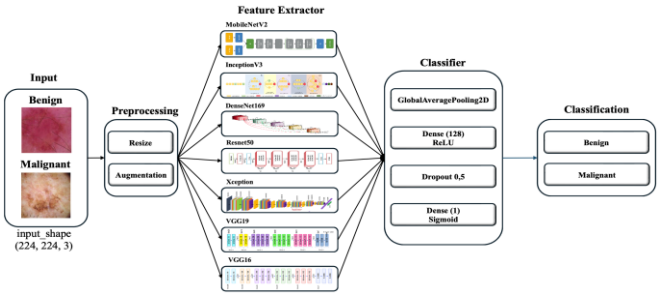


Fig. 3. Proposed model architecture framework.

The initial stage of this research includes collecting skin cancer image data totaling 10,000 images, which are then classified into two main categories, namelyz, benign and malignant. The data is the result of the combination of the three different dataset sources. The first dataset, as many as 3,600 images, was obtained from the Skin Cancer Malignant and Benign Dataset (SCMBD) [29]. Secondly, the source came from the Melanoma Skin Cancer Dataset (MSCD) with 5,000 images [30]. And finally, the third dataset, named the Melanoma Cancer Image Dataset (MCID), consists of 1,400 [31]. All three different datasets are divided into two classes, namely, benign and malignant. Table II shows the details of the data division from each source.

TABLE II. SUMMARY OF THE COMBINED DATASET SOURCE

Class Name	SCMBD Dataset	MSCD Dataset	Additional data MCID	Sub Total
Benign	1,800	2,500	700	5,000
Malignant	1,800	2,500	700	5,000
Total	**3,600**	**5,000**	**1,400**	**10,000**

Training data, validation data, and testing data are the three primary subsets of the skin cancer image data used in this study. To support the experimentation and model evaluation process, two data division schemes are used: the first scheme with a ratio of 60% for training, 20% for validation, and 20% for testing, and an alternative scheme with a ratio of 80%:10%:10%. This division is carried out systematically after all images from the three datasets are combined and standardized in size to 224 × 224 pixels to suit the input needs of the deep learning model [32]. The amount of data in each subset and the division scheme are presented in detail in Table III.

The purpose of using these two division schemes is to ensure that the developed model receives sufficient training data, is optimally validated, and is independently tested so that its performance evaluation is accurate [33]. This approach also allows testing the model's robustness to variations in data proportions and helps find the most effective data sharing configuration in skin cancer image classification. Thus, this data sharing strategy also contributes in reducing evaluation bias and increasing the validity of the overall research results [34].

TABLE III. DATASET DISTRIBUTION

Ratio	Subset	Total data in each subset	Total dataset
Scheme 1 60:20:20	Data training	6,000	10,000
	Data validation	2,000	
	Data testing	2,000	
Scheme 2 80:20:20	Data training	8,000	10,000
	Data validation	1,000	
	Data testing	1,000	

To meet the input dimensions required by all the CNN architectures, the dataset in this study was standardized to 224×224 pixels. To increase the capacity of training data and overcome the limitations of the number of datasets, this study applied an image augmentation technique that was specifically carried out only on a subset of the training data (training set). The augmentation process involves various image transformations, including scaling, rotation, translation, zoom, horizontal flipping, vertical and horizontal position shifts (height and width shift), and brightness adjustment. This augmentation technique not only functions to artificially increase the amount of data but also enriches the variety of visual patterns, which are learned in the training process by the model. With the increasing diversity of augmented data, it is hoped that the model can recognize features more comprehensively and have better generalization capabilities to new data outside the training

dataset [35]. Examples of preprocessing and image augmentation results can be seen in Fig. 4, which shows a visualization of the transformation before the data is used in the training process. Details of the augmentation parameters applied during training are presented in Table IV as a technical reference for implementation, with parameters such as horizontal flip enabled, rotation range of 20 degrees, and zoom, shear, width shift, and height shift range, each valued at 0.2. Meanwhile, Table V shows the amount of augmented data from each dataset division scheme in detail.

TABLE IV. TRAINING DATA AUGMENTATION PARAMETERS

Parameters	Value
Horizontal flip	True
Rotation range	20 degrees
Zoom range	0.2
Shear range	0.2
Width shift range	0.2

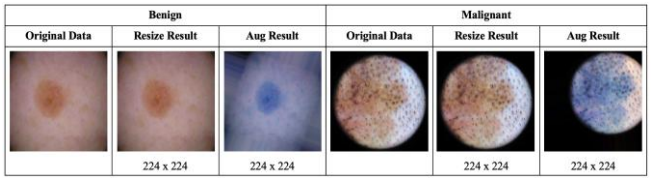


Fig. 4. Example training images from each class.

TABLE V. TRAINING DATASET BEFORE AND AFTER AUGMENTATION

Scheme	Classes	Before augmentation	After augmentation
Scheme 1 60% of dataset	Benign	2,942	5,884
	Malignant	3,058	6,116
Total	**2 classes**	**6,000**	**12,000**
Scheme 2 80% of dataset	Benign	3,998	7,976
	Malignant	4,012	8,024
Total	**2 classes**	**8,000**	**16,000**

Table V shows the amount of data before and after augmentation based on the two dataset division schemes used in this study. In scheme 1, the training data consists of 2,942 benign class images and 3,058 malignant class images. After augmentation, they become 5,884 and 6,116 images, respectively, so that the total training data is 12,000 images. Meanwhile, scheme 2 uses 80% of the entire dataset for training, resulting in 3,998 benign images and 4,012 malignant images, which are multiplied through augmentation to 7,976 and 8,024 images, respectively, with a total of 16,000 training images. The application of this augmentation aims to increase the amount and diversity of data so that the model is better able to recognize visual patterns of both classes of skin cancer accurately.

The training process is carried out using a transfer learning approach using seven CNN architectures, namely MobileNetV2, InceptionV3, Xception, DenseNet169, ResNet50, VGG16, and VGG19. Initial weights that have been previously trained on the ImageNet dataset are initialized on each model, then fine-tuning is carried out so that the model is able to adapt to the visual characteristics of dermatoscopic images of skin cancer [36]. In the training process, the Adam optimization algorithm was used with a learning rate value of 0.001. The performance of each model was monitored periodically based on the accuracy and loss values on the training and validation data in each epoch. The training results were recorded in the form of average training accuracy, average validation accuracy, average training loss, average validation loss, and total training time. After training, the model was saved in an .h5 file and was ready for testing.

The initial stage begins by inputting skin cancer images as input into the model with a size (input shape) of $224 \times 224 \times 3$. The image data consists of two classes, namely benign skin cancer and malignant skin cancer. The training process is carried out on seven CNN model architectures, namely, MobileNetV2, InceptionV3, DenseNet169, ResNet50, Xception, VGG19, and VGG16. The training process is carried out for 20 epochs using the Binary Cross entropy loss function, which is suitable for the binary classification process. A dropout mechanism of 0.5 and an early stopping approach with a 5-epoch patience are used to prevent overfitting. If the model's performance on the validation data does not improve for five consecutive epochs, training will end. A learning rate value of 0.001 as a default of Adam optimizer is also used in the optimization process. At the end of the network (dense layer), 128 neurons are used with the ReLU activation function, and followed by generating the binary output of benign or malignant using the sigmoid activation function in the output layer.

Model evaluation was performed using a subset of test data to measure the final performance of each deep learning architecture in classifying skin cancer images into two categories, namely benign and malignant. Performance evaluation was performed by calculating a number of key classification metrics, such as recall, precision, accuracy, F1-score, and inference time to measure the efficiency of model predictions in the context of real applications [37]. In addition, a confusion matrix was used to summarize the performance of the classification distribution of true (correct) and false (incorrect) predictions from each class, thus offering a more thorough comprehension of the effectiveness of the model in digital image-based classification tasks. The components of the confusion matrix used are True Positive (TP), representing the number of malignant cancer cases that were successfully predicted correctly, while True Negative (TN) shows the number of benign cancer cases that were correctly identified. Meanwhile, False Positive (FP) refers to benign cancer cases that were incorrectly classified as malignant, and False Negative (FN) describes malignant cancer cases that failed to be recognized by the model [37].

The calculation of recall, precision, accuracy, and F1-score follows the formulas outlined in Eq. (1) to Eq. (4). Accuracy provides an overview of the accuracy of the model's predictions. Recall evaluates the model's potential ability to accurately classify cases of malignant cancer, which is very important to avoid the risk of delayed diagnosis. Precision evaluates the accuracy of the model's predictions, avoiding misdiagnosis of healthy patients. Meanwhile, F1-score offers a balanced measure of performance, which particularly relevant when there is an imbalance in the amount of data between classes. It is also known as the harmonic mean of precision and recall. Inference time is also an important indicator in assessing the suitability of a model for application to a real-time detection system in a

clinical context [38]. By applying a combination of these metrics, this study aims to present a thorough evaluation of model performance in digital image-based skin cancer classification.

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}x100 \quad (1)$$

$$\text{Precision}=\frac{TP}{TP+FP}x100 \quad (2)$$

$$\text{Recall}=\frac{TP}{TP+FN}x100 \quad (3)$$

$$\text{F1-score}=2x\frac{\text{Recall}x\text{Precision}}{\text{Recall}+\text{Precision}}x100 \quad (4)$$

IV. Result and Discussion

A. Result

The implementation and comparative evaluation of seven models were carried out using Jupyter Notebook. The Python programming language used in this study is backed by a number of significant libraries, including Keras, NumPy, Pandas, Matplotlib, and TensorFlow, which manage the training and testing stages of the deep learning models. The purpose of this study is to identify the deep learning model that best distinguishes between benign and malignant forms of skin cancer. There are seven types of CNN models tested, namely MobileNetV2, InceptionV3, DenseNet169, ResNet50, Xception, VGG19, and VGG16. All models are trained using an algorithm called Adam, which functions to accelerate and stabilize the training process. Adam was chosen because it is able to adjust its learning rate automatically, so that training is more efficient than other methods, such as RMSprop. A learning rate of 0.001 is often used as an initial value because it is considered stable enough to prevent excessive weight updates but still efficient in reducing loss during training. This value is also very suitable for adaptive optimizers such as Adam, which automatically adjust the learning rate of each parameter. In addition, this learning rate has been proven effective in various studies for models with medium to high complexity and is able to avoid the risk of overshooting the minimum point.

The training process on the seven models was carried out using a fine-tuning strategy, where some early layers of the pre-trained model were kept frozen, while the final layers were retrained using the skin cancer dataset. This strategy allows the model to adjust the general knowledge obtained from large datasets such as ImageNet to the specific characteristics of the dataset used in this study. By applying a dropout rate in the model of 0.5, it is expected to be able to reduce the chance of overfitting while maintaining the ability to capture important details from the training data. Every model has the same initial hyperparameters in order to preserve equality throughout the evaluation process. Table VI shows the results of the training process, including the accuracy, loss, and training time of each model, as a basis for model performance analysis.

Based on the experimental results on seven deep learning architectures with two data splitting scenarios (60:20:20 and 80:10:10), DenseNet169 showed the most outstanding performance. This model managed to achieve the highest validation accuracy, which was 87.07% in the 60:20:20 scenario and increased to 89.57% in the 80:10:10 scenario. In addition, the validation loss value on DenseNet169 was also recorded as the lowest compared to other models, indicating good training stability and generalization ability without symptoms of overfitting. This proves that, despite only the last layer being retrained throughout the fine-tuning procedure, DenseNet169 is able to identify significantly important patterns in the data.

TABLE VI. Comparison of Model Training Result

Model	Split Data	Val Acc (%)	Val Loss (%)
DenseNet169	60:20:20	87.07	29.87
InceptionV3	60:20:20	82.35	39.97
MobileNetV2	60:20:20	83.56	33.63
ResNet50	60:20:20	66.41	64.44
VGG16	60:20:20	81.8	43
VGG19	60:20:20	86.95	31.91
Xception	60:20:20	83.55	38.09
DenseNet169	80:10:10	89.57	23.99
InceptionV3	80:10:10	82.55	36.13
MobileNetV2	80:10:10	85.79	31.29
ResNet50	80:10:10	68.12	61.22
VGG16	80:10:10	84.95	35.82
VGG19	80:10:10	83.89	38.61
Xception	80:10:10	85.69	32.29

Meanwhile, several other models, such as MobileNetV2, Xception, and VGG19, also showed quite good performance with validation accuracy ranging from 83% to 86%, and loss values that are still quite good. These models still have decent potential for use in skin cancer classification, although not as optimal as DenseNet169. On the other hand, ResNet50 consistently recorded the lowest performance, both in terms of validation accuracy and loss rate, in both data scenarios. This indicates that in this study, the ResNet50 architecture is less suitable for the dataset and hyperparameter configuration used. After the training process of the seven deep learning models was completed, each model was saved in .h5 (HDF5) format to preserve the weights and network architecture that had been trained. Saving in this format allows for future reuse of the model without having to repeat the training process from the beginning. The next stage is the testing process for all models using test data that has been previously separated from the training and validation data. The purpose of this testing is to evaluate the final performance of each model on new data that has never been seen before and how well the model can generalize the classification of benign and malignant skin cancers accurately under actual circumstances. The results of testing each model are shown in Table VII.

TABLE VII. COMPARISON OF MODEL TESTING RESULT

Model	Split Data	Test Accuracy (%)	Test Loss (%)	Precision (%)	Recall (%)	F1-score (%)	Inference time
DenseNet169	60:20:20	90.4	9.6	90.4	90.4	90.4	79 s
InceptionV3	60:20:20	86.1	14	86	86.1	86	58 s
MobileNetV2	60:20:20	88	12	88.3	87.7	87.9	16 s
ResNet50	60:20:20	73.7	26.3	74.8	74.2	73.6	55 s
VGG16	60:20:20	85	15	85	85.1	85	194 s
VGG19	60:20:20	88.4	11.6	88.5	88.3	88.4	28 s
Xception	60:20:20	86.7	13.4	86.7	86.5	86.5	78 s
DenseNet169	80:10:10	90.7	9.3	90.7	90.7	90.7	41 s
InceptionV3	80:10:10	83.2	16	84.6	83.2	83.2	25 s
MobileNetV2	80:10:10	87.8	12.2	87.9	87.8	87.8	14 s
ResNet50	80:10:10	72.7	27	72.7	72.7	72.7	40 s
VGG16	80:10:10	82.3	17.7	82.3	82.3	82.3	90 s
VGG19	80:10:10	83.2	16.8	83.2	83.2	83.2	118 s
Xception	80:10:10	86.2	13.8	86.2	86.2	86.2	37 s

Based on the test results, the three best models for skin cancer classification are DenseNet169, MobileNetV2, and VGG19. DenseNet169 shows the best performance with the highest accuracy of 90.7%, as well as equally high recall, precision, and F1-score (90.7%), making it the most reliable and stable model. MobileNetV2 is the most efficient model with the fastest inference time (14 to 16 seconds) and competitive accuracy of up to 88%, suitable for low-resource devices and real-time applications. Meanwhile, VGG19 offers high accuracy (up to 88.4%) and balanced classification metrics, although with a longer inference time (28 to 118 seconds), making it a good choice for systems with high precision requirements and sufficient computing resources.

The analysis results of the three best models based on tests for classifying benign and malignant skin cancer are shown in Fig. 5. A thorough visual comparison of the three models is also provided in Fig. 5 to aid in the study and make it easier to recognize and comprehend the variations in model performance. Based on the results presented in Fig. 5, DenseNet169 (80:10:10) shows the best overall performance with an accuracy value of 90.7%, accompanied by a low loss value (9.3%) and balanced recall, precision, and F1-score scores. The next two best models are VGG19 (60:20:20) and MobileNetV2 (60:20:20), with accuracies of 88.4% and 88.0%, respectively. Considering pure accuracy, the three best models in order are DenseNet169 (80:10:10), VGG19 (60:20:20), and MobileNetV2 (60:20:20). However, if time efficiency is an important factor, MobileNetV2 is worth choosing as a superior alternative model.

B. Discussion

A variety of evaluation metrics of the five models included in this study were analyzed in order to assess the pre-trained CNN models. Several important steps were implemented, including data augmentation to increase the number of samples, and the application of class weighting in the loss function during training to give greater weight to underrepresented classes. This strategy was designed to ensure that the model is able to learn the data without bias towards the majority class. In addition, the balanced batch sampling technique was used to keep the class distribution proportional in each training batch, which helps reduce the impact of class imbalance. The validation and evaluation process was carried out thoroughly using F1-score, recall, precision, and its confusion matrix metrics to ensure prediction accuracy and provide a comprehensive understanding of the strengths and limitations of each CNN model in classifying benign and malignant skin cancers.

Performance optimization was performed on three superior models through the application of fine-tuning techniques. Retraining a number of the model's final layers with a low learning rate and regularization technique reduces the possibility of overfitting. Fine-tuning was applied to DenseNet169, MobileNetV2, and VGG16 by opening some of the final layers, while the initial layers were kept frozen so that the pre-training knowledge was maintained. During the retraining process, the Adam optimizer was used with a standard learning rate value of 0.001 to ensure that the weight update process was stable and efficient.

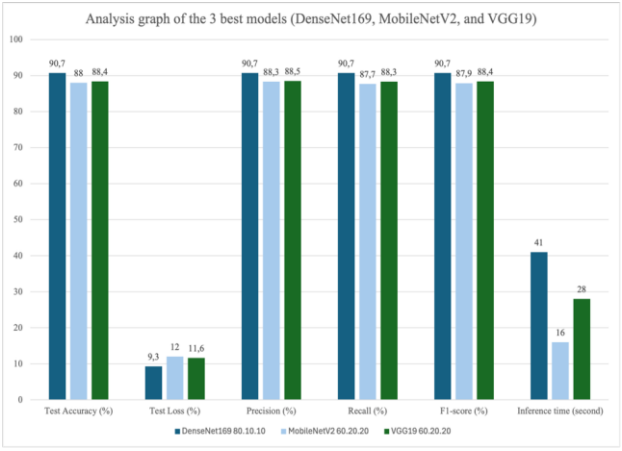


Fig. 5. Analysis graph of the three best models (DenseNet169, MobileNetV2, and VGG19).

This section discusses the best performance obtained from the research results conducted, namely the DenseNet169, MobileNetV2, and VGG19 models. The first result from DenseNet169 produced a test accuracy of 90.7% and a loss of 9.3%. Fig. 6 shows the results with testing data, while Fig. 7 shows the confusion matrix of the DenseNet169 model.

Classification Report:	precision	recall	f1-score	support
benign	0.917	0.893	0.905	495
malignant	0.898	0.921	0.909	505
accuracy			0.907	1000
macro avg	0.907	0.907	0.907	1000
weighted avg	0.907	0.907	0.907	1000

Fig. 6. Classification report of DenseNet169 model.

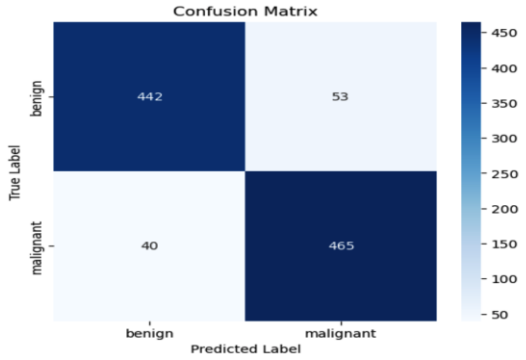


Fig. 7. Confusion matrix of DenseNet169 model.

Based on Fig. 6, the results of the DenseNet169 model show a very good performance in classifying two classes, namely benign and malignant. The precision value, recall, and F1-score of the benign class are 91.7%, 89.3%, and 90.5%, respectively, whereas the malignant class shows 89.8%, 92.1%, and 90.9%. The overall accuracy of the model reaches 90.7%, with macro and weighted average values (macro avg and weighted avg) also consistent at 90.7%, indicating that the model can maintain a balance of performance between classes without significant bias. This reflects how well the model was able to generalize to the test data.

Second, the VGG19 model showed quite good classification performance with an overall accuracy of 88.4%. In the benign class, the model produced a precision value of 87.0% and a recall of 91.2%, with an F1-score reaching 89.1%, which indicates how well the model's ability can identify most benign cases. Meanwhile, in the malignant class, the precision was recorded as higher at 90.1%, but the recall was slightly lower at 85.5%, resulting in an F1-score of 87.7%. This performance reflects that the model tends to be more careful in detecting malignant cases, but is quite effective in reducing false positive predictions. The macro and weighted average values of recall, precision, and F1-score are in the range of 88.3% to 88.6%, respectively, reflecting the stability and balance of classification between classes. Overall, the VGG19 model has good potential for use in image-based skin cancer detection systems with quite competitive performance.

As seen in Fig. 8, it shows the classification report for the VGG19 model, while Fig. 9 shows its confusion matrix. Fig. 8 presents evaluation metrics such as recall, precision, and f1-score for each class (benign and malignant), while Fig. 9 depicts the distribution of correct and incorrect predictions in the form of a confusion matrix. Both figures provide a comprehensive visual representation of the performance of the VGG19 model in classifying skin cancer images.

Classification Report:	precision	recall	f1-score	support
benign	0.870	0.912	0.891	1032
malignant	0.901	0.855	0.877	966
accuracy			0.884	1998
macro avg	0.886	0.883	0.884	1998
weighted avg	0.885	0.884	0.884	1998

Fig. 8. Classification report of the VGG19 model.

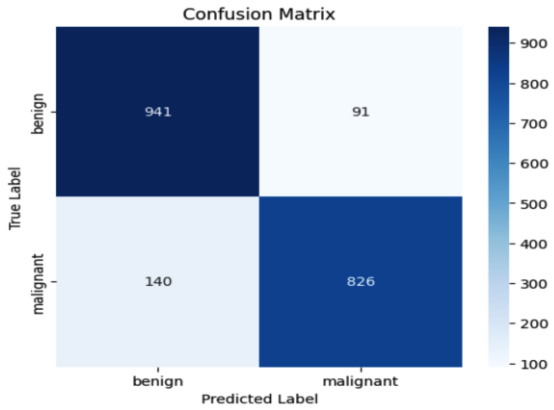


Fig. 9. Confusion matrix of the VGG19 model.

And at last, MobileNetV2, produces a test accuracy of 88% and a loss of 12%. The results of the MobileNetV2 model with testing data and its confusion matrix can be seen in Fig. 10 and Fig. 11, respectively. The MobileNetV2 model shows good performance in classifying benign and malignant categories, with an overall accuracy of 88.0%. The benign class has a precision value of 85.7%, a recall of 92.6%, and an F1-score of 89.0%, indicating that it is quite efficient in identifying most benign cases, although there are several false predictions. In contrast, in the malignant class, the precision value is recorded as higher, namely 91.0%, but with a lower recall of 82.9%, indicating that although the positive predictions for this class are quite accurate, the model still misses a number of malignant cases. The macro and weighted average values of recall, precision, and F1-score range from 87.8% to 88.4%, reflecting a relatively balanced model performance between classes. Despite some disparity in detection between classes, the model still shows good potential for use in general skin cancer image classification.

Based on the evaluation results of the three deep learning models tested, it can be concluded that DenseNet169 is the model with the best performance in skin cancer image classification. With a test accuracy of 90.7% and a loss value of only 9.3%, this model shows a very good balance between recall, precision, and F1-score for both classes, namely benign and malignant. The high macro and weighted average values of 90.7% indicate that this model is able to fairly classify both classes without significant bias. These results indicate that DenseNet169 is not only accurate but also stable and reliable in handling data variations, making it the main choice for a digital image-based skin cancer classification system.

```
Classification Report:
              precision    recall  f1-score   support

      benign       0.857     0.926     0.890      1051
   malignant       0.910     0.829     0.868       949

    accuracy                         0.880      2000
   macro avg       0.884     0.878     0.879      2000
weighted avg       0.882     0.880     0.880      2000
```

Fig. 10. Classification report of MobileNetV2 model.

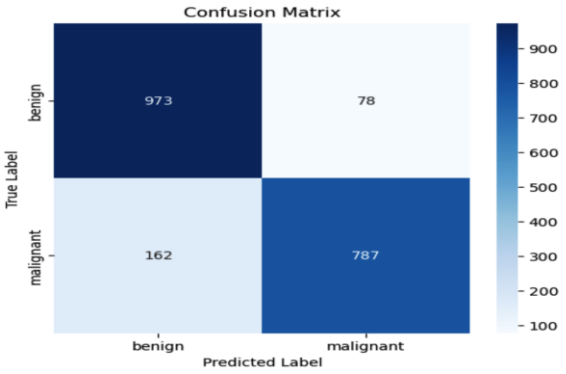


Fig. 11. Confusion matrix of MobileNetV2 model.

The VGG19 and MobileNetV2 models also demonstrate competitive performance, achieving accuracies of 88.4% and 88.0%, respectively. VGG19 exhibits higher recall for the benign class, while MobileNetV2 attains superior precision for the malignant class. Although a minor recall imbalance exists between the two classes, both models deliver stable and consistent performance. MobileNetV2, in particular, shows an advantage in inference efficiency, making it suitable for deployment on resource-constrained devices. Overall, all three models exhibit strong potential for integration into computer-aided skin cancer diagnostic systems. DenseNet169 remains the most robust and accurate architecture, while MobileNetV2 offers a lightweight yet effective alternative for real-time or embedded applications. While this study primarily focuses on classification accuracy, future research may explore explainable AI (XAI) methods such as Grad-CAM to visualize feature attention on dermoscopic images, thereby improving model interpretability and clinical trust. To further validate the proposed approach, Table VIII compares the results of this study with those reported in previous works.

TABLE VIII. COMPARISON OF MODEL EVALUATION RESULTS WITH PREVIOUS STUDIES

Proposed method	Acc (%) This Study	Acc (%) [19]	Acc (%) [20]	Acc (%) [21]	Acc (%) [22]	Acc (%) [23]
Densenet169	90.7	89.7	-	-	-	-
VGG19	88.4	-	-	-	87	88
MobileNetv2	88	-	85.6	85	-	-

V. CONCLUSION

This study evaluated three deep learning architectures, namely DenseNet169, MobileNetV2, and VGG19, for automatic skin cancer classification using digital dermatoscopic images. The dataset combined three public Kaggle sources containing 10,000 benign and malignant cases, evaluated under two data-split schemes (60:20:20 and 80:10:10). Among the models, DenseNet169 achieved the highest accuracy of 90.7%, demonstrating superior robustness and balanced precision-recall performance. MobileNetV2 and VGG19 also showed competitive accuracy (88.0 to 88.4%) with consistent reliability, where MobileNetV2 excelled in inference efficiency, making it suitable for mobile or real-time clinical applications.

These findings confirm the potential of deep learning-based approaches to enhance early skin cancer detection. Each model offers distinct advantages that can be adapted to various healthcare environments depending on system requirements and resource availability. However, this study is limited by dataset diversity in terms of ethnicity, lighting, and geographic distribution, and it does not yet incorporate advanced ensemble learning, lesion segmentation, or explainable AI techniques. Future work will focus on improving generalization through multi-source data integration, exploring model interpretability via XAI methods such as Grad-CAM, and developing lightweight web- or mobile-based diagnostic tools for real-world deployment.

Overall, the results demonstrate that CNN-based architectures, particularly DenseNet169, can significantly contribute to advancing medical image analysis and AI-assisted dermatological diagnostics. In clinical practice, such models can accelerate the diagnostic process, reduce misclassification risks, and improve early detection outcomes. Rather than replacing dermatologists, these systems are designed to support clinical decision-making, helping to improve patient recovery rates and the overall quality of healthcare services.

ACKNOWLEDGMENT

This research was funded by a research grant from the Ministry of Education, Science, and Technology of the Republic of Indonesia. The authors would like to express their sincere gratitude to the Ministry for the financial support that made this study possible. The authors also thank N. Khasanah for her valuable assistance during the research implementation, and L. Kurniawati for her contributions to preparing and organizing the manuscript. Special appreciation is also extended to E. Firasari and E. H. Hermaliani for their help in finalizing the study for publication.

REFERENCES

[1] A. H. Roky et al., "Overview of skin cancer types and prevalence rates across continents," Cancer Pathogenesis and Therapy, vol. 3, no. 2, pp. 89–100, Mar. 2025, doi: 10.1016/j.cpt.2024.08.002.

[2] A. Kurva, M. Korikani, V. Mohan, and R. K. Kancha, "Skin Cancer," in Biomedical Aspects of Solid Cancers, Singapore: Springer Nature Singapore, 2024, pp. 235–252. doi: 10.1007/978-981-97-1802-3_21.

[3] Z. G. Attal et al., "Advanced and Metastatic Non-Melanoma Skin Cancer. Epidemiology, Risk Factors, Clinical Features, and Treatment Options," Biomedicines, vol. 12, no. 7, p. 1448, Jun. 2024, doi: 10.3390/biomedicines12071448.

[4] S. Arige, L. R. Atmakuri, R. Shaik, M. Gude, V. K. Ghanta, and R. Alluri, "Digital Dermoscopy: Advancements in Skin Cancer Diagnosis and Monitoring," Biomedical and Pharmacology Journal, vol. 18, no. December Spl Edition, pp. 33–43, Jan. 2025, doi: 10.13005/bpj/3071.

[5] A. El Mrabet, M. Benaly, I. Alihamidi, B. Kouach, L. Hlou, and R. El Gouri, "Enhancing Early Detection of Skin Cancer in Clinical Practice with Hybrid Deep Learning Models," Engineering, Technology & Applied Science Research, vol. 15, no. 2, pp. 20927–20933, Apr. 2025, doi: 10.48084/etasr.9753.

[6] S. Haque, F. Ahmad, V. Singh, D. M. Mathkor, and A. Babegi, "Skin Cancer Detection Using Deep Learning Approaches," Cancer Biother Radiopharm, vol. 40, no. 5, pp. 301–312, Jun. 2025, doi: 10.1089/cbr.2024.0161.

[7] N. Merlina, A. Prasetio, I. Zuniarti, N. A. Mayangky, D. N. Sulistyowati, and F. Aziz, "Deep CNN Models for Detecting Cervical Cancer in Pap Smear Images," TEM Journal, pp. 1073–1083, May 2025, doi: 10.18421/TEM142-09.

[8] N. Merlina, A. Prasetio, I. Zuniarti, N. Almira Mayangky, D. Nur Sulistyowati, and F. Aziz, "Improving Early Detection of Cervical Cancer Through Deep Learning-Based Pap Smear Image Classification," Journal of Applied Data Sciences, vol. 6, no. 2, pp. 952–968, May 2025, doi: 10.47738/jads.v6i2.576.

[9] E. Firasari, N. Khasanah, F. L. D. Cahyanti, D. N. Kholifah, U. Khultsum, and F. Sarasati, "Performance Evaluation of ResNet50 and MobileNetV2 in Skin Cancer Image Classification with Various Optimizers," in 2024 International Conference on Information Technology Research and Innovation (ICITRI), IEEE, Sep. 2024, pp. 376–380. doi: 10.1109/ICITRI62858.2024.10698943.

[10] N. Khasanah and M. N. Winnarto, "Application of Deep Learning with ResNet50 for Early Detection of Melanoma Skin Cancer," Journal Medical Informatics Technology, pp. 16–20, Mar. 2024, doi: 10.37034/medinftech.v2i1.31.

[11] Daniati Uki Eka Saputri, Nurul Khasanah, F. Aziz, and Taopik Hidayat, "Enhancing Skin Cancer Classification Using Optimized InceptionV3 Model," Journal Medical Informatics Technology, pp. 65–69, Sep. 2023, doi: 10.37034/medinftech.v1i3.14.

[12] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," Artif Intell Rev, vol. 57, no. 4, p. 99, Mar. 2024, doi: 10.1007/s10462-024-10721-6.

[13] J. A. AYENI, "Convolutional Neural Network (CNN): The architecture and applications," Applied Journal of Physical Science, vol. 4, no. 4, pp. 42–50, Dec. 2022, doi: 10.31248/AJPS2022.085.

[14] M. M. Musthafa, M. T R, V. K. V, and S. Guluwadi, "Enhanced skin cancer diagnosis using optimized CNN architecture and checkpoints for automated dermatological lesion classification," BMC Med Imaging, vol. 24, no. 1, p. 201, Aug. 2024, doi: 10.1186/s12880-024-01356-8.

[15] C. Kavitha, S. Priyanka, M. P. Kumar, and V. Kusuma, "Skin Cancer Detection and Classification using Deep Learning Techniques," Procedia Comput Sci, vol. 235, pp. 2793–2802, 2024, doi: 10.1016/j.procs.2024.04.264.

[16] K. Nawaz et al., "Skin cancer detection using dermoscopic images with convolutional neural network," Sci Rep, vol. 15, no. 1, p. 7252, Mar. 2025, doi: 10.1038/s41598-025-91446-6.

[17] B. Ozdemir and I. Pacal, "A robust deep learning framework for multiclass skin cancer classification," Sci Rep, vol. 15, no. 1, p. 4938, Feb. 2025, doi: 10.1038/s41598-025-89230-7.

[18] D. S. Hieu, D. T. Phuc, L. N. Ton, T. C. Hung, and N. C. Cuong, "Improving Architectures of VGG16, AlexNet, and ResNet50 Models for Skin Cancer Classification," 2025, pp. 562–571. doi: 10.1007/978-3-031-90194-2_39.

[19] R. Pathania and P. Behki, "Skin Cancer Detection Using Deep Learning," in 2024 Sixth International Conference on Computational Intelligence and Communication Technologies (CCICT), IEEE, Apr. 2024, pp. 568–575. doi: 10.1109/CCICT62777.2024.00095.

[20] O. Şahin and M. S. Yıldırım, "Performance Comparison of Deep Learning Architectures for Skin Cancer Classification," in 2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP), IEEE, Sep. 2024, pp. 1–8. doi: 10.1109/IDAP64064.2024.10710839.

[21] D. Moturi, R. K. Surapaneni, and V. S. G. Avanigadda, "Developing an efficient method for melanoma detection using CNN techniques," J Egypt Natl Canc Inst, vol. 36, no. 1, p. 6, Feb. 2024, doi: 10.1186/s43046-024-00210-w.

[22] I. Ahmad, B. S. Alsulami, and F. Alqurashi, "Enhancing Skin Cancer Detection with Transfer Learning and Vision Transformers," International Journal of Advanced Computer Science and Applications, vol. 15, no. 10, 2024, doi: 10.14569/IJACSA.2024.01510104.

[23] I. A. Kandhro et al., "Performance evaluation of E-VGG19 model: Enhancing real-time skin cancer detection and classification," Heliyon, vol. 10, no. 10, p. e31488, May 2024, doi: 10.1016/j.heliyon.2024.e31488.

[24] S. Mushtaq and O. Singh, "A deep learning based architecture for multi-class skin cancer classification," Multimed Tools Appl, vol. 83, no. 39, pp. 87105–87127, Jul. 2024, doi: 10.1007/s11042-024-19817-1.

[25] V. Anand, S. Gupta, S. R. Nayak, D. Koundal, D. Prakash, and K. D. Verma, "An automated deep learning models for classification of skin disease using Dermoscopy images: a comprehensive study," Multimed Tools Appl, vol. 81, no. 26, pp. 37379–37401, Nov. 2022, doi: 10.1007/s11042-021-11628-y.

[26] D. Albashish, N. Almansour, A. Abdullah, H. M. J. Mustafa, M. R. AlSayyed, and O. Alrashdan, "Design an Ensemble Pretrained Deep Learning Model for Classification of Melanoma Skin Cancer Images," in 2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA), 2025, pp. 1–7.

[27] N. R. Pratama, D. R. I. M. Setiadi, I. Harkespan, and A. A. Ojugo, "Feature Fusion with Albumentation for Enhancing Monkeypox Detection Using Deep Learning Models," Journal of Computing Theories and Applications, vol. 2, no. 3, pp. 427–440, Feb. 2025, doi: 10.62411/jcta.12255.

[28] S. Matiray and L. K. Singh, "Deep Learning for Early Skin Cancer Detection: A Comparative Study on Hybrid CNN Models," 2025. doi: 10.3233/ATDE250013.

[29] C. Fanconi, "Skin Cancer: Malignant vs. Benign," Kaggle.com. Accessed: Jun. 11, 2025. [Online]. Available: https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign.

[30] M. Hasnain Javid, "Melanoma Skin Cancer Dataset of 10000 Images," Kaggle.com. Accessed: Jun. 11, 2025. [Online]. Available: https://www.kaggle.com/datasets/hasnainjaved/melanoma-skin-cancer-dataset-of-10000-images.

[31] B. Mittal, "Melanoma Cancer Image Dataset," Kaggle.com. Accessed: Jun. 11, 2025. [Online]. Available: https://www.kaggle.com/datasets/bhaveshmittal/melanoma-cancer-dataset.

[32] A. Mahbod, N. Saeidi, S. Hatamikia, and R. Woitek, "Evaluating pre-trained convolutional neural networks and foundation models as feature extractors for content-based medical image retrieval," Eng Appl Artif Intell, vol. 150, p. 110571, Jun. 2025, doi: 10.1016/j.engappai.2025.110571.

[33] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," International Journal of Advanced Computer Science and Applications, vol. 15, no. 2, 2024, doi: 10.14569/IJACSA.2024.0150235.

[34] P. Georgiadis, E. V. Gkouvrikos, E. Vrochidou, T. Kalampokas, and G. A. Papakostas, "Building Better Deep Learning Models Through Dataset Fusion: A Case Study in Skin Cancer Classification with Hyperdatasets," Diagnostics, vol. 15, no. 3, p. 352, Feb. 2025, doi: 10.3390/diagnostics15030352.

[35] Y. Zhong, W. Zhou, and Z. Wang, "A Survey of Data Augmentation in Domain Generalization," Neural Process Lett, vol. 57, no. 2, p. 34, Mar. 2025, doi: 10.1007/s11063-025-11747-9.

[36] A. Yilmaz, G. Gencoglan, R. Varol, A. A. Demircali, M. Keshavarz, and H. Uvet, "MobileSkin: Classification of Skin Lesion Images Acquired Using Mobile Phone-Attached Hand-Held Dermoscopes," J Clin Med, vol. 11, no. 17, p. 5102, Aug. 2022, doi: 10.3390/jcm11175102.

[37] S. Sathyanarayanan, "Confusion Matrix-Based Performance Evaluation Metrics," African Journal of Biomedical Research, pp. 4023–4031, Nov. 2024, doi: 10.53555/AJBR.v27i4S.4345.

[38] H. E. C. da Silva et al., "The use of artificial intelligence tools in cancer detection compared to the traditional diagnostic imaging methods: An overview of the systematic reviews," PLoS One, vol. 18, no. 10, p. e0292063, Oct. 2023, doi: 10.1371/journal.pone.0292063.

TabNet–XGBoost Hybrid Model for Student Performance Prediction and Customized Feedback

Anupama Prasanth
College of Computer Studies, University of Technology Bahrain, Salmabad, Kingdom of Bahrain

***Abstract*—Virtual Learning Environments (VLEs) have emerged as a cornerstone of modern education, enabling large-scale delivery of learning materials, assessments, and interactions in fully or partially online formats. The dynamic and self-paced nature of VLEs makes the early prediction of learner scores crucial for timely intervention and support. The existing frameworks either underperform in capturing complex, non-linear relationships in heterogeneous educational data or lack interpretability mechanisms necessary for actionable interventions. This study proposes a TabNet–XGBoost hybrid model with SHAP-based interpretability for score range classification in VLE contexts, using the Open University Learning Analytics Dataset (OULAD). Data preprocessing involved cleaning, encoding, normalization, feature engineering, and score band derivation, producing an enriched feature matrix integrating demographic, assessment, and engagement indicators. TabNet's sequential attentive feature selection extracted a latent representation of the most informative variables, which was subsequently refined by XGBoost to produce sharper decision boundaries for four distinct score ranges. SHAP values were computed post-prediction to identify domain-specific performance drivers, enabling alignment with a structured feedback module across seven predefined learning domains. Experimental results demonstrated a classification accuracy of 98.8% on the test set, outperforming the baseline frameworks. The SHAP-driven feedback mechanism provided interpretable, domain-targeted insights, enhancing the model's practical applicability for educators and academic support teams. By integrating high predictive accuracy with transparent reasoning and actionable feedback, the proposed framework addresses both the technical and pedagogical requirements of early performance prediction in online learning environments, offering a scalable solution for real-time academic monitoring and intervention.**

Keywords*—*Virtual learning environments; student performance prediction; TabNet; XGBoost; SHAP; feedback generation; quality education

I. Introduction

Virtual Learning Environments have transformed the educational landscape by enabling interactive, resource-rich, and quality education platforms where learners can access course materials, engage with peers, and receive instructional guidance regardless of physical location [1]. These digital ecosystems offer unprecedented flexibility, as students can customize learning pace and learning paths and integrate academic activities into diverse personal and professional schedules [2]. VLEs provide scalability for institutions, support diverse learning modalities, and facilitate continuous assessment opportunities. Despite these advantages, learner performance within such environments can vary significantly, influenced by factors such as digital engagement habits, self-directed learning skills, and adaptability to technology-mediated instruction [3]. A central challenge remains the consistent and accurate evaluation of learner progress, alongside the timely provision of constructive feedback [4]. Without proactive performance monitoring and targeted intervention, learners in VLEs may experience reduced motivation, disengagement, or even withdrawal from the course.

Traditional approaches for performance assessment in VLEs often rely on periodic quizzes, assignment submissions, and final examinations. While these methods are administratively convenient and familiar to both educators and learners, they are predominantly retrospective, identifying performance gaps only after significant time has elapsed. Such delay limits the scope for early corrective action and diminishes the effectiveness of feedback, particularly in fast-paced or highly competitive academic contexts [5]. Moreover, these approaches tend to prioritize academic scores over other formative indicators such as participation in discussions, interaction with learning resources, or the quality of peer collaboration, all of which are vital for a holistic understanding of learner progress in virtual environments.

The adoption of advanced analytics in VLEs, with machine learning (ML) and deep learning (DL) algorithms, predictive analytics, and learning dashboards, enables educators to track and forecast student performance using a wide array of behavioral and academic indicators [6]. However, existing implementations often fall short in adaptability, failing to dynamically adjust to evolving learner behaviors over time [7]. Some models focus narrowly on quantitative academic metrics, neglecting qualitative engagement data, while others operate as opaque "black boxes", limiting their interpretability and diminishing trust among educators and stakeholders. This underscores the need for an integrated, transparent, and adaptive performance prediction framework tailored for VLEs. An ideal system should deliver early, accurate forecasts of learner performance, informed by both academic and behavioral data, while ensuring that feedback is specific, actionable, and timely.

The primary objective and questions addressed in this study are mentioned below:

- Development of a hybrid TabNet–XGBoost Framework that integrates TabNet's attention-driven feature selection with XGBoost's gradient-boosted decision refinement, optimized for score range classification in VLEs.

- Implementation of an interpretable feedback generation module that maps SHAP-based feature importance values to seven predefined learning domains, enabling personalized, domain-specific recommendations for academic performance improvement.

The succeeding sections of the study are organized as follows: Section II offers a comprehensive analysis of the latest advances in student performance prediction in VLEs and highlights the current research constraints. Section III delineates the proposed methodology. The experimental results are highlighted in Section IV along with a comprehensive assessment of the model's functionality. Finally, Section V completes the research by summarizing the main conclusions and highlighting the possible domains for future research.

II. Related Work

Liu et al. [8] suggested the use of temporal engagement data collected from online learning platforms for student performance prediction. The OULAD, consisting of clickstream data from 5,341 distance learners across 12 different VLEs, was the source of the study. Two primary feature sets were designed based on the weekly and monthly aggregation of click counts, treating them as structured panel data. A number of classifiers, including Logistic Regression (LR), k-Nearest Neighbors (KNN), Random Forest (RF), Gradient Boosted Trees, 1D Convolutional Neural Networks (1D-CNN) and Long Short-Term Memory networks (LSTM) were compared in the study. LSTM outperformed others with a maximum accuracy of 90.25% and interactions with the homepage, subpages, content, and quizzes were identified as the most predictive behaviors in the feature importance analysis. As the framework failed to incorporate assessment performance or generate personalized feedback, its applicability in student score improvement was limited.

Arashpour et al. [9] proposed hybridized ML models for student exam performance prediction utilizing OULAD. The hybrid models integrated Support Vector Machines (SVM) and Artificial Neural Networks (ANN) with a Teaching–Learning-Based Optimizer (TLBO) for both classification and regression. TLBO performed feature selection and optimized the ANN structure in parallel, identifying the optimal subset of predictive input variables. The SVM model, with eight selected features, achieved a classification accuracy of 86.10%, while the ANN model achieved 84.94% on the test data. For regression, Support Vector Regression (SVR) outperformed ANN, achieving a correlation coefficient $R > 0.7$. Engagement, measured through clickstream behavior and performance in ongoing assessments, was identified as the most influential predictor. Heavy dependence on numerical engagement metrics, lacking semantic interpretation or personalized feedback mechanisms, hampered the practical use of the framework.

Rao et al. [10] proposed DL models for early prediction of student academic performance utilizing log data obtained from academic systems. The dataset comprised log records from 108 students enrolled in an "Information Science" course was utilized which captured sequential interactions with the learning environment. The architecture employed a Recurrent Neural Network (RNN) that processed sequential student interaction logs, capturing temporal dependencies in learning behavior patterns over the course timeline. The framework dynamically updated the internal state with each input sequence, enabling it to learn correlations between past engagement and future academic outcomes. The RNN framework with 84.3% accuracy outperformed the Decision Tree classifier and ensemble learning approaches. As there was no integration of time-based engagement trends or automated feedback generation, the deployment of the framework for continuous formative assessment and personalized academic support was constrained.

Jawad et al. [11] proposed a Random Forest classifier, enhanced by the SMOTE oversampling technique, to predict student academic performance in VLE. The study utilized OULAD and constructed dynamic student profiles across six-time intervals (120 to 260 days), integrating engagement patterns and assessment scores. The framework was retrained per time segment to simulate real-time academic monitoring, and the integration of SMOTE mitigated the impact of imbalanced class distributions. The model achieved its highest testing accuracy of 84.2% at 260 days, suggesting that richer late-term data boosts predictive strength. Additional benchmarking revealed that XGBoost and Logistic Regression achieved accuracies of 84.3% and 80.9%, respectively. The reliance on separate models across multiple time intervals introduced structural redundancy and limited scalability for real-time deployment.

Yu et al. [12] proposed a recurrent neural network (RNN) for student learning outcome prediction in an online Artificial Intelligence course, focusing on identifying at-risk students using a limited number of commonly available LMS features. A number of DL models: Simple Recurrent Network (SRN), LSTM, Gated Recurrent Unit (GRU), Multilayer Perceptron (MLP), and CNN, as well as ML frameworks: LR, SVM, RF, and decision trees (DT), were compared. The SRN and CNN models achieved the highest overall accuracy (93%), while LSTM and GRU followed closely with 91% and 92%, respectively. Further week-wise performance analysis revealed that GRU showed the best predictive accuracy (87%) by the 18th week, followed by SRN (86%) and CNN (84%), outperforming all classical ML models. The rigid framework for all prediction weeks overlooked varying learning dynamics at different stages of the course, hampering adaptability to fluctuating student engagement patterns.

Kusumawardani and Alfarozi [13] proposed a transformer encoder-based DL framework for early prediction of student performance using log activity data from learning management systems (LMS). The study utilized the OULAD and was designed for both daily and weekly prediction tasks for the early identification of at-risk students. The transformer architecture was fine-tuned by evaluating the effects of components of positional encoding, feature aggregation strategies and weighted loss functions. The framework performed best without positional encoding, and weekly feature aggregation yielded higher accuracy. For the withdrawn vs. pass-distinction task, an accuracy of 83.17% was observed at just 20% of the course duration, improving further to 90% accuracy by the end. Across all tasks, the transformer outperformed LSTM by 1-3% in accuracy and 3-7% in F1-score. The model complexity and resource-intensive architecture hindered deployment in real-time or low-resource educational environments.

Chen et al. [14] proposed Explainable Student Performance Prediction (ESPP) for early detection of at-risk learners in VLEs. The data for the study was collected from a Digital Transformation course at Gadjah Mada University, Indonesia, comprising weekly student activity logs during online instruction. The hybrid deep learning framework integrated CNN and LSTM networks, alongside the hybrid SMOTE oversampling technique to balance class distributions. CNN layers extracted spatial patterns from temporal activity sequences, which were then passed through LSTM layers to model time-dependent engagement trends. Conv-LSTM model achieved the highest accuracy of 91%, followed by CNN-LSTM at 88%, outperforming LSTM (85%), SVM (80%) and LR (64%) on evaluation in the six-week. However, the computational complexity was a major drawback, hindering scalability or real-time deployment in low-resource educational platforms.

Adnan et al. [15] suggested a predictive model to identify at-risk students in VLEs, MOOCs, and LMS platforms using OULAD. A number of multiple ML and DL algorithms were evaluated, including RF, DT, SVM, LR, KNN, MLP, ANN, and Naive Bayes (NB). The study incorporated key variables such as assessment scores, clickstream data indicative of engagement intensity, and temporal indicators [15]. On evaluation, the RF classifier performed consistently across different stages of course progression and achieved 79% accuracy at 20% course completion, which further improved to 88% at 60% and peaked at 91% at 100% of the course length. The high dependence on feature engineering that did not generalize well across different online learning environments without significant adaptation constrained the potential of the study.

Riestra-González et al. [16] proposed a course-agnostic framework to predict student performance from LMS log files in the early stages of course delivery. The study utilized data from 5,112 courses hosted on Moodle at the University of Oviedo, involving over 29,000 students across diverse disciplines. Multiple classification models were employed, such as DT, NB, LR, SVM, and MLP, to detect at-risk, failing, and exceptional students at different course phases. The MLP model achieved the highest accuracy, ranging from 80.1% at 10% of course length to 90.1% at 50%, while DT followed closely with accuracies ranging from 79.5% to 89.6% over the same intervals. A clustering technique was also applied to identify six consistent student interaction patterns, four of which were identified to be strongly correlated with student performance: early answering of quizzes, prompt viewing of LMS resources, early viewing of course assignments, and procrastination in viewing course content, which indicated a higher risk of failure. The reliance on static interaction patterns, without incorporating the temporal sequence or evolution of student behaviors over time, hampered the generalizability of the framework.

Yağcı [17] proposed a machine learning-based predictive framework for predicting students' final exam grades using minimal yet impactful academic indicators. The dataset for the study comprised midterm grades, faculty affiliation, and departmental information of 1,854 undergraduate students enrolled in the Turkish Language-I course at a public university in Turkey. Six ML models: RF, SVM, LR, NB, KNN, and Neural Network (NN) were evaluated in the study. The framework extracted predictive insights from numerical academic features without incorporating behavioral or interaction data. On evaluation, RF and NN achieved the highest classification accuracy, both reaching 74.6%, followed by SVM with 73.5%, LR at 71.7%, NB at 71.3% and kNN at 69.9%. The sole focus on academic variables without the incorporation of behavioral or engagement-related features limited the study.

Chen et al. [18] proposed an Attention-Based ANN (Attn-ANN) model for early prediction of at-risk students by integrating attention mechanisms across both time and feature dimensions. Data for the study were collected from the M2B system at Kyushu University, which included LMS-based records such as attendance, report submissions, and course access. The framework employed dual attention layers: on the temporal dimension to identify important weeks and on the feature dimension to prioritize learning activities. The attention weights are integrated directly into a standard artificial neural network, enabling it to dynamically adjust the influence of each time step and feature during the learning process. The Attn-ANN achieved an accuracy of 64.3% in the Programming Techniques (PT) course and up to 89.5% in the Digital Signal Processing (DSP) course, outperforming conventional models like MLP, LSTM, and GRU in both early and progressive weeks. The sensitivity of attention weight calibration required manual adjustment or retraining across different course structures and educational settings.

Hakkal and Ait Lahcen [19] suggested integration of XGBoost with logistic regression-based models for learner performance prediction within Intelligent Tutoring Systems (ITS). Three regression models were utilized: Item Response Theory (IRT), Performance Factor Analysis (PFA), and DAS3H. Eight real-world datasets from varied sources, including four ASSISTments skill-builder math datasets, two KDD Cup Algebra datasets, the Statics engineering dataset, and a new Moodle-Morocco dataset, were employed in the study. The framework analyzed historical ITS interaction logs and estimated the probability of a learner answering future questions correctly. The XGBoost-enhanced PFA outperformed standard PFA in seven datasets, DAS3H also improved on the ASSISTments17 dataset, while IRT's performance remained stable across datasets, indicating less benefit from XGBoost integration. The highest accuracy value of 84.9% was observed for DAS3H-LR on Bridge-Algebra06, followed by 84.3% by PFA-XGBOOST on the Bridge-Algebra06 dataset. The worst accuracy of 68.1% was exhibited by IRT-XGBOOST on the Assistments09 dataset. However, XGBoost consumed a long execution time for large datasets and required advanced hyperparameter tuning and specialized feature encoding, hampering the scalability in real-world scenarios.

Iatrellis et al. [20] proposed a two-phase machine learning approach combining unsupervised and supervised learning in student outcome prediction. Data from the Computer Science Department at the University of Thessaly, Greece, was utilized in the study to forecast degree completion time and the likelihood of student enrollment. The first phase employed the K-Means algorithm to cluster students based on educational factors and metrics, which identified three coherent student groups. The prediction models were developed for each cluster in the second phase for customized predictions. The clustering-

guided model outperformed non-clustering models, achieving an accuracy of 80.5%, compared to 75.5% in the non-clustering approach. As the student grouping was solely from a data-driven perspective, the faculty insights or academic advisor evaluations were overlooked in the study, hampering the content relevance of the predicted outcomes.

A. Research Gap

Despite significant advancements in predictive modeling for student performance, several research gaps persist. Many existing approaches, including DL architectures like BiLSTM, CNN, and LSTM, as well as ML models such as RF, SVM, and hybrid optimization-based methods, have demonstrated high accuracy using datasets such as OULAD, LMS logs, or Intelligent Tutoring Systems [8] [14] [22]. However, a recurring limitation is the narrow scope of input features; most frameworks rely heavily on clickstream data, assessment scores, or basic academic records, overlooking rich contextual, socio-cultural, and behavioral attributes that could improve model generalization [15] [17]. Furthermore, temporal modeling has often been static or rigid, failing to adapt to evolving engagement patterns across course stages [12]. While certain methods incorporate early prediction capabilities, few integrate automated personalized feedback or intervention strategies to translate predictions into actionable academic support [10]. Scalability is also a challenge; resource-intensive architectures like transformers and CNN-LSTM struggle in low-resource educational settings, while approaches with segmented time-based retraining introduce redundancy. Lastly, class imbalance handling techniques like SMOTE improve accuracy but may reduce real-world applicability where such balancing is infeasible, highlighting the need for models that maintain robustness in naturally imbalanced datasets [11]. This calls for a novel model that could predict the student performance in VLEs with high accuracy and tailor feedback for academic improvements while considering both academic and behavioral attributes.

III. Materials and Methods

The proposed study integrates TabNet and XGBoost for enhanced student performance prediction in VLEs. The hybrid architecture leverages TabNet's sequential decision-step feature selection and representation learning capabilities, coupled with XGBoost's robust non-linear modeling, to generate precise score range classifications. The OULAD serves as the primary data source and SHAP was employed for the customized feedback generation. Fig. 1 represents the basic architecture of the proposed model.

A. Dataset Description

The proposed study employs the publicly available Open University Learning Analytics Dataset (OULAD) from Kaggle, a comprehensive real-world dataset sourced from the VLE of the Open University, the largest distance-learning institution in the United Kingdom [21]. The dataset comprises records from seven distinct modules delivered across multiple academic years, with presentations denoted for the first and second semesters. It integrates three major data categories: student demographic attributes, assessment-related data, and VLE interaction logs. The student demographic attributes include age band, gender, geographic location, disability status, prior education level, and Index of Multiple Deprivation (IMD) band.

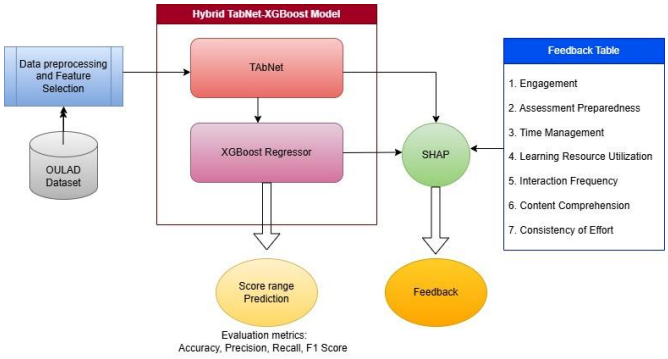


Fig. 1. Basic architecture of the proposed model.

The assessment-related data contains marks, submission dates, and completion status for assignments, quizzes, and final examinations. The VLE interaction logs include time-stamped clickstream data capturing student engagement patterns, including content access, forum participation, and quiz activity. The dataset's multimodal nature, combining demographic, behavioral, and performance indicators, aligns closely with the proposed model architecture, enabling both sequential learning through historical interactions and explainability through feature-level attribution. Given its richness, scale, and diversity, OULAD provides an ideal foundation for developing high-accuracy predictive models with actionable, domain-specific feedback to improve student performance outcomes

B. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is performed for the systematic examination of the structural composition and statistical properties of the dataset, enabling the identification of underlying patterns, distributions, and relationships between variables. The process is crucial for the detection of anomalies, missing values, and potential data imbalances, while also highlighting key behavioral and demographic trends relevant to student performance. Insights from the EDA informed both engineering decisions and the subsequent design of the predictive modelling framework.

The distribution of results across individual modules illustrated in Fig. 2, reveals substantial variation in student performance patterns, offering critical insights for targeted academic interventions.

As depicted, certain modules, such as BBB and FFF (as named in the dataset), exhibit the highest overall enrolments and are characterized by a comparatively large proportion of withdrawn students, suggesting potential structural or delivery-related challenges. Conversely, modules like AAA and EEE display lower participation volumes but notable failure rates, which may indicate concentrated difficulties among smaller cohorts. The presence of distinctions is relatively modest across all modules, with only minor variation between courses, while pass rates remain the most frequent positive outcome. This module-level stratification provides a more nuanced understanding than aggregate performance statistics, enabling the identification of courses where tailored pedagogical

strategies or enhanced learning support may yield the most significant improvements. Such granularity aligns directly with the proposed model's capacity to generate module-specific performance predictions and customized feedback, thereby facilitating informed decision-making for educators and administrators aiming to reduce attrition and improve academic success rates.

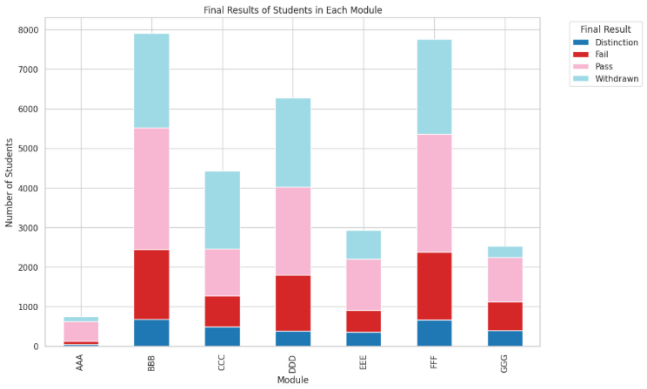


Fig. 2. Final results of students in each module.

Fig. 3 represents the distribution of key learning activity features, segmented by students' final results.

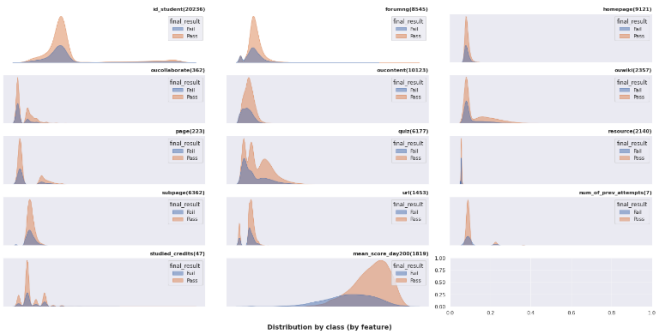


Fig. 3. Distribution of key learning activity features.

Each subplot corresponds to a specific feature from the OULAD dataset with kernel density estimations illustrating the relative frequency of values for each outcome class. Across most activity metrics, students who passed tend to show higher engagement levels, as indicated by a broader spread and higher density in the upper value ranges compared to their failing counterparts. Notably, features like quiz and mean score day demonstrate a clear separation between the two classes, indicating strong predictive value. Conversely, some variables, such as oucollaborate and ouwiki, exhibit overlapping distributions, implying limited discriminative power. The skewed patterns in the number of previous attempts and studied credits suggest that prior academic history and course load may influence performance outcomes.

Fig. 4 compares the average submission dates between B Semesters and J Semesters (for the second and first semesters, respectively) expressed in days from the start of the academic term. The results show that on average, students in the second semesters submit their work earlier compared to those in the first semesters. This difference of nearly 10 days suggests possible variations in course scheduling, assessment deadlines or student pacing between the two semester types. Such temporal patterns can inform predictive modeling by highlighting semester-specific behaviors that may influence engagement and performance outcomes.

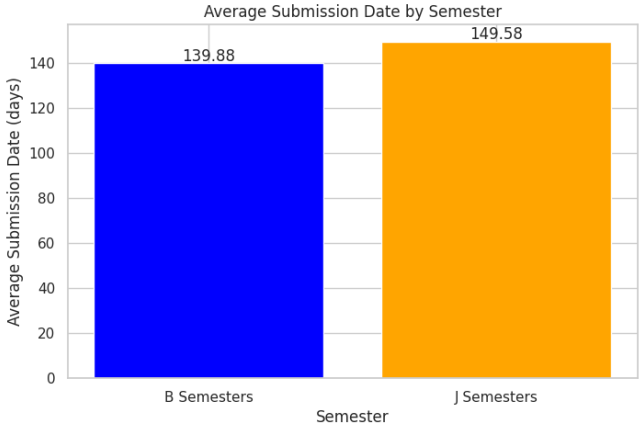


Fig. 4. Average submission date by semester.

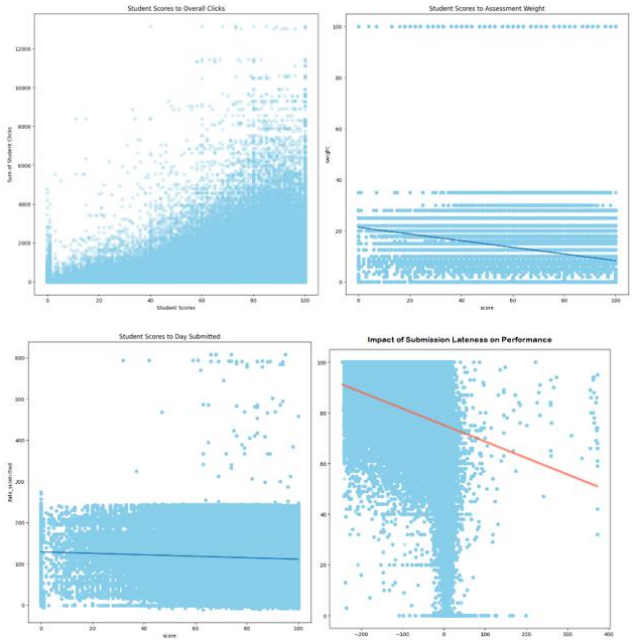


Fig. 5. Student score analysis.

Fig. 5 presents scatter plot analyses exploring the relationships between student scores and multiple engagement or temporal features in the dataset. The Student Scores to Overall Clicks reveal a positive association between total VLE clicks and student scores, suggesting that higher engagement levels are generally linked to better academic performance. The Student Scores to Assessment Weight indicates that most assessments have lower weightings, with a dense concentration of points near zero weight and minimal variation in scores across higher weights. The Student Scores to Day Submitted shows a weak negative trend, suggesting that later submission dates may be modestly associated with lower scores, though the relationship appears minimal. The Impact of Submission

Lateness on Performance illustrates a clearer negative relationship, where increased lateness correlates with lower scores, as captured by the downward-sloping regression line. Collectively, these plots highlight that engagement intensity is a strong positive predictor, while lateness is a notable negative factor influencing student performance, whereas assessment weight and submission timing exhibit weaker or more diffuse relationships.

C. Data Preprocessing

The data preprocessing is a critical stage in predictive modeling, ensuring that raw educational data is transformed into a structured, consistent, and analytically suitable format. This process involves cleaning and standardizing heterogeneous data sources, handling missing or inconsistent values, and encoding variables to align with the input requirements of advanced deep learning architectures.

Data cleaning is performed first to remove noise, inconsistencies, and missing values from the dataset. For the raw dataset $= \{x_i, y_i\}^N_{i=1}$, where x_i is the feature vector and y_i, the corresponding label for the i^{th} student, the missing values in a numerical feature f_j are replaced using mean imputation, as in Eq. (1):

$$f_j^* = \frac{\sum_{i=1}^{N} f_{ij}}{N_j^{obs}} \quad (1)$$

where, N_j^{obs} represents the number of observations without missing values for feature f_j, and f_j^* is the imputed value. For categorical variables, the missing entries were replaced by the mode, as in Eq. (2):

$$f_j^* = arg \underset{c \in C_j}{max\ count\ (c)} \quad (2)$$

where, C_j is the set of possible categories for f_j. Outliers in continuous variables are detected using the z-score method, as in Eq. (3):

$$z_{ij} = \frac{f_{ij} - \mu_j}{\sigma_j} \quad (3)$$

where, μ_j and σ_j denote the mean and standard deviation of feature f_j, respectively. Instances with $|z_{ij}| > 3$ are flagged for removal or transformation. Following data cleaning, feature engineering is performed to transform the dataset into a structured form compatible with the model. Categorical variables such as gender, region, and highest education were encoded using one-hot encoding, producing binary indicator vectors for each observation. For a categorical feature f_j with K_j distinct categories, each observation i is represented, as in Eq. (4):

$$f_{i,j}^{(k)} = \begin{cases} 1\ ;if\ observation\ i\ belongs\ to\ category\ k \\ 0\ ;otherwise \end{cases} \forall k \in [1, K_j] \quad (4)$$

Numerical attributes, including click counts, assessment scores, and time-on-task, were normalized to a uniform scale to ensure balanced gradient updates during model training. The min–max transformation is applied, as in Eq. (5):

$$f'_{i,j} = \frac{f_{i,j} - min\ (f_j)}{max(f_j) - min\ (f_j)} \quad (5)$$

where, $f'_{i,j}$ is the normalized value for feature f_j in observation i. To capture engagement dynamics, temporal VLE logs were aggregated into weekly indicators of student activity. The weekly engagement score for student i in week w was computed, as in Eq. (6):

$$E_{i,w} = \frac{\sum_{t \in w} clicks_{i,t}}{\Delta t_w} \quad (6)$$

where, $clicks_{i,t}$ is the total VLE interactions at timestamp t and Δt_w is the number of active days in week w. In addition, assessment-related features were derived by computing the weighted average of scores, as in Eq. (7):

$$A_{(i)} = \frac{\sum_{k=1}^{K} w_k \cdot m_{(i,k)}}{\sum_{k=1}^{K} w_k} \quad (7)$$

where, $m_{(i,k)}$ denotes the mark of student i in assessment k and w_k is its respective weight. Temporal score change was measured, as in Eq. (8), capturing improvement or decline between consecutive assessments.

$$\Delta A_{(i,t)} = m_{(i,t)} - m_{(i,t-1)} \quad (8)$$

The final engineered feature matrix X^* integrates demographic encodings, normalized continuous attributes, weekly engagement profiles, and assessment-based indicators, providing a comprehensive, multi-view representation of each learner. For the score range prediction task, continuous marks were discretized into performance bands to align with the feedback generation module. The score band for student i is as in Eq. (9):

$$Band_i = \begin{cases} 0\ ;if\ 0 \le S_i < 40 \\ 1\ ;if\ 40 \le S_i < 60 \\ 2\ ;if\ 60 \le S_i < 80 \\ 3\ ;if\ 80 \le S_i \le 100 \end{cases} \quad (9)$$

where, S_i denotes the final computed score from all available assessments. These engineered features form the input vector X_i for subsequent representation learning in the proposed framework. Feature selection is further performed to retain critical informative predictors and reduce model complexity. The mutual information criterion is applied between each feature f_j and the target variable y, as in Eq. (10):

$$MI(f_j, y) = \sum_{f_j} \sum_{y} p(f_j, y)\ log \frac{p(f_j, y)}{p(f_j)\ p(y)} \quad (10)$$

where, $p(f_j, y)$ is the joint probability distribution of feature f_j and target y and $p(f_j), p(y)$ are the respective marginal probabilities. Features with $MI(f_j, y) < \tau$ are discarded, where τ is the data-driven threshold. The processed dataset X^* is then split into training and testing subsets in the ratio 80:20. Class imbalance handling is applied only to the training set to prevent information leakage. Using the Synthetic Minority Oversampling Technique (SMOTE), new synthetic samples are generated, as in Eq. (11):

$$x_{new} = x_i + \lambda \cdot (x_{nn} - x_i) \quad (11)$$

where, x_i is a minority class sample, x_{nn} is one of its KNN and $\lambda \in [0,1]$ is a random interpolation factor. Finally, the balanced training set is structured for the proposed architecture, with each observation X_i represented, as in Eq. (12):

$$X_i = [X_i^{demo}, X_i^{eng}, X_i^{asses}, X_i^{temp}] \quad (12)$$

where, X_i^{demo} contains demographic encodings, X_i^{eng} with weekly engagement scores, assessment-derived metrics in X_i^{asses} and X_i^{temp} contains temporal change indicators to facilitate multi-view input compatibility with the hierarchical attention encoder.

D. Model Deployment

1) TabNet architecture: TabNet is a DL architecture designed specifically for tabular data, combining the interpretability of DTs with the representational power of NNs [24]. The model processes data in multiple sequential decision steps, where at each step a subset of the most relevant features is selected through an attentive feature-masking mechanism. The process begins with an input vector $X_i \in \mathbb{R}^d$ for student i, which is first normalized and passed through a shared feature transformer network $FT_{shared}(\cdot)$. The initial hidden representation is as in Eq. (13):

$$H_i^{(0)} = FT_{shared}(X_i; W_f) \quad (13)$$

where, W_f are learnable weights of the shared feature transformer block. At each decision step $t \in \{1,2,\dots,T\}$, TabNet computes an attention mask $M^{(t)}$ over the features using an Attentive Transformer. The mask is modulated by a prior scale vector $P^{(t)} \in [0,1]^d$ which controls feature reuse and is initialized as $P^{(t)} = 1$ for $t = 1$. The mask is given as in Eq. (14):

$$M_i^{(t)} = sparsemax(P^{(t)} \odot AT^{(t)}(H_i^{(t-1)}; W_a)) \quad (14)$$

where, W_a is the attention weight matrix, $\odot$ denotes element-wise multiplication, and the sparsemax activation ensures that only a subset of features receive non-zero weights, enhancing interpretability. The masked feature vector is obtained as in Eq. (15) and passed through the decision step network as in Eq. (16):

$$\tilde{H}_i^{(t)} = M_i^{(t)} \odot H_i^{(t-1)} \quad (15)$$

$$(D_i^{(t)}, H_i^{(t)}) = DecisionBlock^{(t)}(\tilde{H}_i^{(t)}; W_d) \quad (16)$$

where, $D_i^{(t)} \in \mathbb{R}^k$ is the decision output contributing directly to the prediction and $H_i^{(t)}$ is the transformed feature representation for the current step. The prior scale vector for the next step is updated to reduce the weight of features already selected, allowing partial reuse via the relaxation parameter $\gamma > 1$ as in Eq. (17):

$$P^{(t+1)} = P^{(t)} \odot (\gamma - M_i^{(t)}) \quad (17)$$

The final prediction is the sum of decision outputs across all T steps as in Eq. (18):

$$\hat{y}_i = \sum_{t=1}^{T} D_i^{(t)} \quad (18)$$

In the proposed framework, however, instead of using only $\hat{y}_i$ from the aggregated decision outputs, a latent representation is formed by concatenating the transformed hidden vectors from all decision steps as in Eq. (19):

$$Z_i = Concat(H^{(1)}, H^{(2)}, \dots, H^{(T)}) \quad (19)$$

The enriched feature vector Z_i is then fed into the XGBoost model that refines the prediction of the student's score range and enables sharper decision boundaries. The basic architecture of TabNet is depicted in Fig. 6.

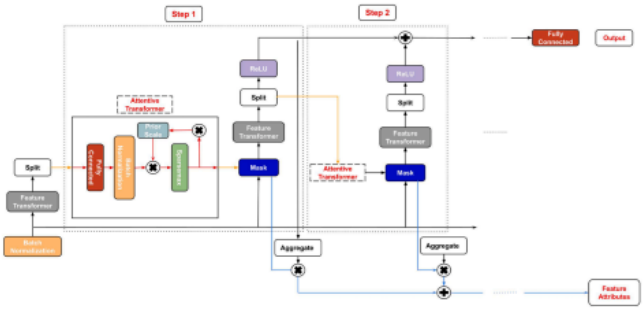


Fig. 6. Basic architecture of TabNet.

2) XGBoost Regressor: Extreme Gradient Boosting (XGBoost) is an optimized implementation of the gradient boosting framework designed for scalability, efficiency, and high predictive accuracy on structured data [23]. The algorithm builds an ensemble of decision trees sequentially, where each new tree f_t is trained to minimize the residual errors of the previous trees. For the dataset $\{(X_i, y_i)\}_{i=1}^{n}$, the model prediction at iteration t is as in Eq. (20):

$$\hat{y}_i = \hat{y}_i^{(t-1)} + \eta f_t(Z_i) \quad (20)$$

where, η is the learning rate and f_t represents the freshly added regression tree. The optimization process minimizes regularized objective function, as shown in Eq. (21):

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t) \quad (21)$$

where, $l(y_i, \hat{y}_i)$ is the differentiable loss function and $\Omega(f_t)$ is the regularization term defined, as shown in Eq. (22):

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \quad (22)$$

where, T is the number of leaves in tree f_t, w_j is the weight assigned to leaf j, γ penalizes the creation of excessive leaves to control the model complexity, and λ controls L2 regularization on leaf weights.

Fig. 7 represents the basic architecture of XGBoost improves traditional gradient boosting through features such as parallelized tree construction, sparsity-aware split finding and weighted quantile sketch for efficient handling of missing values. Its ability to model complex, non-linear relationships makes it highly effective for educational datasets where interactions between demographic, engagement and assessment features can strongly influence performance predictions.

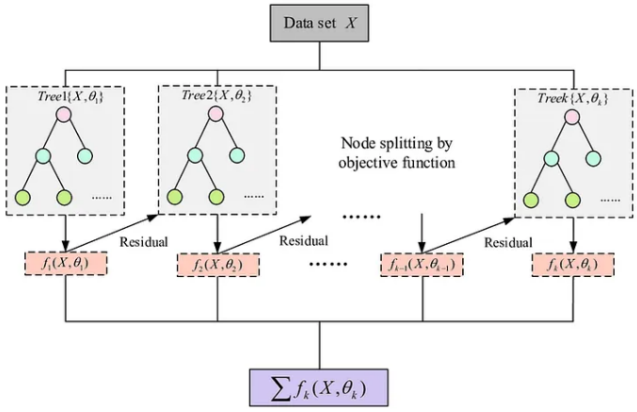


Fig. 7. Basic architecture of XGBoost Regressor.

3) Proposed TabNet-XGBoost hybrid model: In the proposed TabNet-XGBoost architecture, TabNet serves as the representation learning module, extracting a compact yet information-rich latent feature vector from the pre-processed student data. The output of the TabNet module, the concatenated vector Z_i, captures both global and stepwise feature importance patterns, effectively encoding demographic, engagement and assessment-related information into a single high-dimensional embedding. This enriched feature vector Z_i is then passed into the XGBoost Regressor, which acts as the final prediction layer. In regression mode, XGBoost iteratively builds trees to refine the score prediction. After convergence, the continuous score prediction is discretized into four score bands representing performance ranges. This integration allows TabNet to perform interpretable and sparsity-driven feature selection, while XGBoost captures non-linear relationships and sharpens class separation boundaries. The combination results in a robust hybrid model capable of predicting students' score ranges with high precision while retaining interpretability for SHAP-based feedback generation in later stages.

Following score band prediction, the model incorporates SHapley Additive exPlanations (SHAP) to ensure interpretability and to support the customized feedback generation module. SHAP, grounded in cooperative game theory, assigns each feature a Shapley value ϕ_j representing its average marginal contribution to the model's prediction across all possible feature subsets. For a given student i with a feature set F and model prediction $f(X_i)$ the Shapley value for feature j s computed as in Eq. (23):

$$\phi_j = \sum_{S \subseteq F\{j\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!} \left[f_{S\cup\{j\}}(X_i) - f_s(X_i)\right] \quad (23)$$

where, S is a subset of features excluding j, $f_{S\cup\{j\}}(X_i)$ is the model output when j is included and $f_s(X_i)$ is the output without j. This ensures a fair and coherent measure of feature importance, regardless of ordering or correlation. In the proposed hybrid framework, SHAP is applied after the XGBoost stage, taking the enriched latent vector Z_i as the feature input and the XGBoost model's predicted score range as the target for explanation [24]. Following SHAP analysis, the feature importance values for each individual student are mapped onto seven pre-defined feedback domains: Engagement, Assessment Preparedness, Time Management, Learning Resource Utilization, Interaction Frequency, Content Comprehension and Consistency of Effort. For each domain d, the mean SHAP value $s_{i,d}$ is computed by averaging SHAP contributions of features belonging to that domain as in Eq. (24):

$$s_{i,d} = \frac{1}{|F_d|}\sum_{f \in F_d} SHAP_{i,f} \quad (24)$$

where, F_d is the set of features assigned to the domain d. Domains with high positive SHAP values indicate areas that most contributed to predicted high performance, while high negative SHAP values signal weaknesses requiring intervention. A feedback table is then generated for each student, with rows representing the domains and columns indicating performance status along with the recommendations. This integration ensures that the system not only predicts student performance with high accuracy but also delivers interpretable, actionable and structured recommendations to improve learning outcomes. The algorithm for the proposed model is as shown below (see Algorithm 1).

Algorithm 1: TabNet-XGBoost hybrid model for Student performance prediction and customized feedback

Input:

- *OULAD dataset* $D = \{(X_i, y_i)\}_{i=1}^{N}$
- *Y: Label vector corresponding to X*
- *Label vector* $\in \{0,1,2,3\}$*, each value for different score range*

Output:

- *Predicted class label* $Y \in \{0,1,2,3\}$ *and customized feedback*

Begin:

Data collection

- *Load OULAD dataset*
- *Extract feature matrix X and label vector Y*

Data Preprocessing and Feature extraction

- *Missing values imputation:*

$$f_j^* = \frac{\sum_{i=1}^{N} f_{ij}}{N_j^{obs}}$$

- *Outliner detection and handling:*

$$z_{ij} = \frac{f_{ij} - \mu_j}{\sigma_j}, remove\ if\ |z_{ij}| > 3$$

- *One hot encoding for categorical features:*

$$f_{i,j}^{(k)} = \begin{cases} 1\,; if\ observation\ i\ belongs\ to\ category\ k \\ 0\,; otherwise \end{cases}$$

- *Normalization:*

$$f'_{i,j} = \frac{f_{i,j} - \min(f_j)}{\max(f_j) - \min(f_j)}$$

- *Various score computations*

Weekly engagement score, $E_{i,w} = \frac{\sum_{t \in w} clicks_{i,t}}{\Delta t_w}$

Weighted Assessment Score, $A_{(i)} = \frac{\sum_{k=1}^{K} w_k \cdot m_{(i,k)}}{\sum_{k=1}^{K} w_k}$

Temporal score change, $\Delta A_{(i,t)} = m_{(i,t)} - m_{(i,t-1)}$

- *Discretize score into performance bands*

$$Band_i = \begin{cases} 0\,; if\ 0 \le S_i < 40 \\ 1\,; if\ 40 \le S_i < 60 \\ 2\,; if\ 60 \le S_i < 80 \\ 3\,; if\ 80 \le S_i \le 100 \end{cases}$$

- *Train-Test Split: Split into 80:20 ratio* $X_{train}, X_{test}, Y_{train}, Y_{test} = train_test(X, Y, test_{size} = 0.2)$

TabNet

- *Transform input:* $H_i^{(0)} = FT_{shared}(X_i; W_f)$
- *For each decision step t:*
  a. *Compute attention mask,* $M_i^{(t)} = sparsemax\,(P^{(t)} \odot AT^{(t)}(H_i^{(t-1)}; W_a)$
  b. *Apply mask and process through decision block,* $(D_i^{(t)}, H_i^{(t)}) = DecisionBlock^{(t)}(\widetilde{H}_i^{(t)}; W_d)$
  c. *Update the prior scale,* $P^{(t+1)} = P^{(t)} \odot (\gamma - M_i^{(t)})$
- *Concatenate hidden states:* $Z_i = Concat(H^{(1)}, H^{(2)}, \ldots, H^{(T)})$

XGBoost prediction

- *Initialize Predictions,* $\hat{y}_i^{(0)} = 0$
- *For* $t = 1\ to\ T$
  a. *Fit tree* f_t *to residuals*
  b. *Update prediction:* $\hat{y}_i = \hat{y}_i^{(t-1)} + \eta f_t(Z_i)$
- *Optimize regularized objective*

$$L(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t)$$

SHAP based interpretation

- *Apply SHAP to* Z_i *and* $\hat{y}_i$ *to compute per dpomain feature importance values*

$$\phi_j = \sum_{S \subseteq F\{j\}} \frac{|S|!\ (|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{j\}}(X_i) - f_s(X_i) \right]$$

Feedback Generation

- *For each domain d, map the SHAP score to feedback category*

$$s_{i,d} = \frac{1}{|F_d|} \sum_{f \in F_d} SHAP_{i,f}$$

Model Compilation and Training

- *Compile model with loss = sparse categorical crossentropy, learning rate = 0.001, optimizer = Adam, Epochs =50*
- *Train model: model.fit (X_train, y_train)*

Evaluation and Model Saving

- *Evaluate model: model. evaluate (X_test, Y_test)*
- *Tune hyperparameters*
- *Save the model*

End

E. Simulation Setup

The proposed TabNet-XGBoost hybrid model was implemented using a high-performance computational environment. The system configuration included an Intel Core i7 processor, an NVIDIA GeForce GTX 1080Ti GPU and 32 GB of RAM that collectively ensured efficient handling of the intensive training and evaluation processes involved in IoT security anomaly identification and classification. To develop the model, Keras API built on TensorFlow and python was selected as the programming language. Google Colaboratory (Colab) was used for model training and testing, taking advantage of its free access to powerful GPUs and cloud-based execution environment that improved the study's accessibility and reproducibility. The hyperparameters that affect the model behavior significantly are manually selected before training and have a direct impact on the framework's rate of convergence, generalization capability and ultimate classification performance. The full list of hyperparameters and training settings employed in the study is summarized in Table I.

TABLE I. HYPERPARAMETER SPECIFICATIONS

Hyperparameters	Values
Epochs	50
Dropout	0.2
Activation function	ReLU
Optimizer	ADAM
Loss function	Sparse categorical cross entropy
Batch size	32
Learning Rate	0.001

IV. RESULTS AND DISCUSSION

A set of standard evaluation metrics has been employed to evaluate the performance of the proposed model, as illustrated in Eq. (25) to Eq. (28). These measures are mathematically computed using the core elements of the confusion matrix: True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). Accuracy indicates the overall correctness while recall and precision highlight the framework's effectiveness in the prediction of score ranges without many misses or false alarms.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (25)$$

$$Precision = \frac{TP}{TP+FP} \quad (26)$$

$$Recall = \frac{TP}{TP+FN} \quad (27)$$

$$F1 - score = 2 \times \frac{precision \times Recall}{Precision + Recall} \quad (28)$$

Accuracy visualizes a model's learning progress across training epochs, showing how well the model is fitting the data. The accuracy plot indicates improvements in predictive performance over time, while the loss plot reflects the model's error reduction. Comparing training and validation curves helps identify overfitting, underfitting or stable convergence. These insights guide hyperparameter tuning, regularization adjustments and architecture refinements to improve model generalization and performance.

The accuracy plot in Fig. 8 demonstrates a consistent improvement in both training and validation accuracy over the 50 epochs, starting from around 0.90 and 0.87, respectively. Both curves rise steadily, with validation accuracy closely tracking training accuracy, peaking at approximately 0.989 for training and 0.988 for validation.

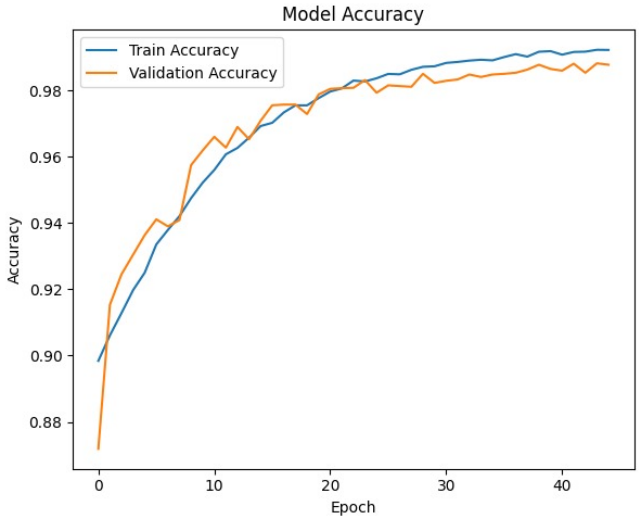


Fig. 8. Accuracy plot.

The minimal gap between the curves suggests that the hybrid architecture generalizes well to unseen data and refrains from significant overfitting. This strong and parallel upward trend indicates that the chosen architecture and hyperparameters are effective in progressively improving classification performance throughout training.

Fig. 9 illustrates the confusion matrix that indicates the proposed model's remarkable classification performance across all four score bands. Misclassifications are minimal and primarily occur between adjacent score bands, such as 0–Less than 40 and 40–Less than 60, suggesting occasional boundary overlap in predicted scores. Overall, the distribution reflects both high accuracy and stability in multi-class categorization.

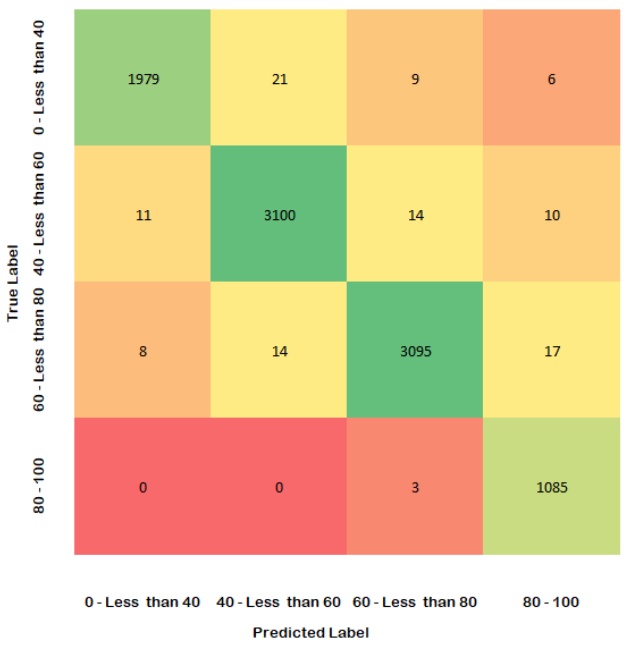


Fig. 9. Confusion matrix.

The metric evaluation results in Fig. 10 demonstrate the robustness and effectiveness of the TabNet-XGBoost hybrid model across all performance indicators.

An accuracy of 98.79%, indicated that the vast majority of predictions correctly matched the actual class labels across the four score bands. The precision value of 98.54% reflects the model's ability to minimize FPs, guaranteeing that most of the predicted instances for each class were indeed correct. The recall score of 98.89% shows the correct identification of true instances across all classes, with very few false negatives.

The F1-score of 98.71%, harmonically balances precision and recall, confirms the framework's consistent performance in both detecting true cases and avoiding misclassifications. Collectively, these findings highlight that the TabNet-XGBoost architecture maintains a balanced trade-off between precision and recall while achieving excellent overall classification performance.

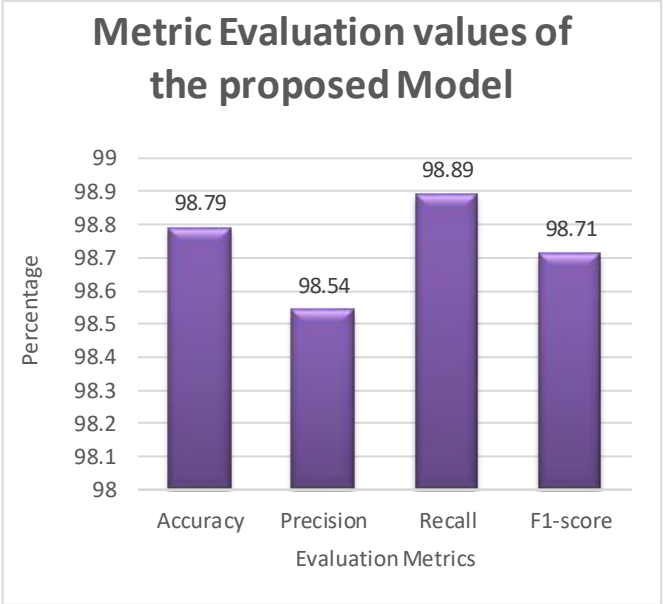


Fig. 10. Metric evaluation values.

Deep learning architectures such as LSTM, RNN and Conv-LSTM improve temporal dependency modeling, achieving accuracy levels exceeding 90%, but these models tend to suffer from high computational costs, longer training times and limited interpretability. Transformer-based and attention-augmented models address feature importance explicitly, yet still struggle with optimizing for structured tabular data, where sparsity and heterogeneous feature types prevail. In contrast, the proposed TabNet–XGBoost hybrid model leverages TabNet's interpretable feature selection capabilities with XGBoost's powerful non-linear decision boundaries, achieving an accuracy of 98.8%. This not only surpasses the performance of prior models but also offers enhanced interpretability and adaptability to varied educational datasets, making it a robust choice for real-world deployment. Fig. 11 represents graphical representation of the accuracy comparison.

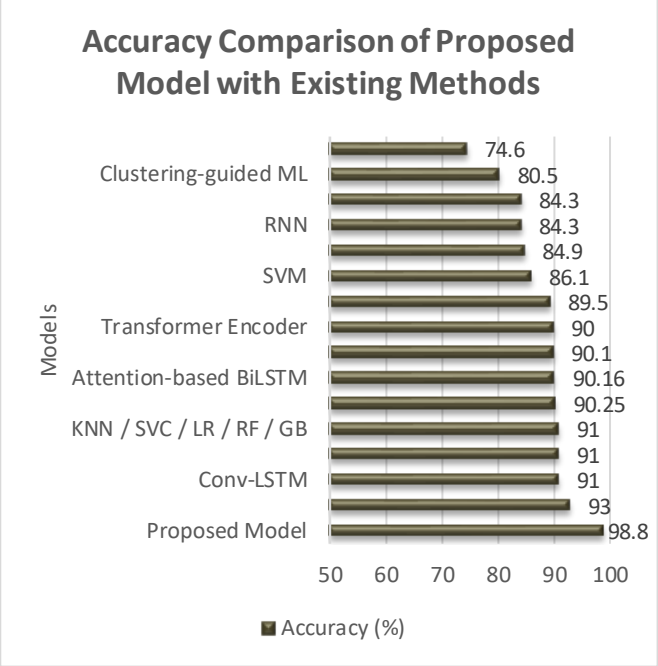


Fig. 11. Accuracy comparison.

V. Conclusion

This study presented a hybrid TabNet–XGBoost framework for predicting student performance and providing targeted feedback, demonstrating the potential of combining deep learning-based feature selection with powerful gradient boosting techniques for educational data mining. By integrating TabNet's attentive feature-masking mechanism, the model efficiently identified and utilized the most informative attributes from the OULAD, while XGBoost refined the latent feature representations to enhance score range classification accuracy The incorporation of SHAP enabled model interpretability by quantifying the contribution of each feature to individual predictions, facilitating the development of a feedback module tailored to seven pedagogical domains. The proposed architecture achieved an accuracy of 98.8%, significantly outperforming conventional ML and DL baselines. Furthermore, the interpretability offered by SHAP ensured that predictions were not only highly accurate but also actionable, aligning predictive analytics and instructional decision-making. Future works may focus on expanding the feature space to incorporate additional behavioral, social and temporal engagement indicators, as well as validating the framework across multiple institutions.

References

[1] Caprara, L., & Caprara, C. (2022). Effects of virtual learning environments: A scoping review of literature. *Education and information technologies*, *27*(3), 3683-3722.

[2] Turan, Z., Kucuk, S., & Cilligol Karabey, S. (2022). The university students' self-regulated effort, flexibility and satisfaction in distance education. *International Journal of Educational Technology in Higher Education*, *19*(1), 35.

[3] Caraig, R. V., & Fabro, R. G. (2022). Students' engagement and study habits in the online distance learning environment. *Int. J. Sci. Res. in Multidisciplinary Studies Vol*, *8*(9).

[4] Carby, N. (2023). Personalized feedback in a virtual learning environment. *Journal of Educational Supervision*, *6*(1), 36.

[5] Almalki, A. D. A., & Mohammed, A. I. (2022). The Effect of Immediate and Delayed Feedback in Virtual Classes on Mathematics Students' Higher Order Thinking Skills. *Journal of Positive School Psychology*, *6*(6).

[6] Ismanto, E., Ghani, H. A., & Saleh, N. I. B. M. (2023). A comparative study of machine learning algorithms for virtual learning environment performance prediction. *IAES International Journal of Artificial Intelligence*.

[7] Torkhani[1], W., & Rezgui, K. (2025, February). OULAD MOOC Student Performance Prediction using Machine and Deep Learning. In *Proceedings of International Conference on Decision Aid and Artificial Intelligence (ICODAI 2024)* (Vol. 12, p. 228). Springer Nature.

[8] Liu, Y., Fan, S., Xu, S., Sajjanhar, A., Yeom, S., & Wei, Y. (2022). Predicting student performance using clickstream data and machine learning. *Education Sciences*, *13*(1), 17.

[9] Arashpour, M., Golafshani, E. M., Parthiban, R., Lamborn, J., Kashani, A., Li, H., & Farzanehfar, P. (2023). Predicting individual learning performance using machine-learning hybridized with the teaching-learning-based optimization. *Computer Applications in Engineering Education*, *31*(1), 83-99.

[10] Rao, G. M., & Kumar, P. K. K. (2021). Students performance prediction in online courses using machine learning algorithms. *United Int. J. Res. Technol*, *2*(11), 74-79.

[11] Jawad, K., Shah, M. A., & Tahir, M. (2022). Students' academic performance and engagement prediction in a virtual learning environment using random forest with data balancing. *Sustainability*, *14*(22), 14795.

[12] Yu, Chih-Chang & Wu, Leon. (2021). Early Warning System for Online STEM Learning—A Slimmer Approach Using Recurrent Neural Networks. Sustainability. 13. 12461. 10.3390/su132212461.

[13] Kusumawardani, S. S., & Alfarozi, S. A. I. (2023). Transformer encoder model for sequential prediction of student performance based on their log activities. *Ieee Access*, *11*, 18960-18971.

[14] Chen, H. C., Prasetyo, E., Tseng, S. S., Putra, K. T., Kusumawardani, S. S., & Weng, C. E. (2022). Week-wise student performance early prediction in virtual learning environment using a deep explainable artificial intelligence. *Applied Sciences*, *12*(4), 1885.

[15] Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., ... & Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. *Ieee Access*, *9*, 7519-7539.

[16] 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.

[17] Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11.

[18] Chen, Y., Wei, G., Liu, J., Chen, Y., Zheng, Q., Tian, F., ... & Wu, Y. (2023). A prediction model of student performance based on self-attention mechanism. *Knowledge and Information Systems*, *65*(2), 733-758.

[19] Hakkal, S., & Ait Lahcen, A. (2024). XGBoost to enhance learner performance prediction. *Computers and Education: Artificial Intelligence*, *7*, 100254.

[20] Iatrellis, O., Savvas, I. K., Fitsilis, P., & Gerogiannis, V. C. (2021). A two-phase machine learning approach for predicting student outcomes. *Education and Information Technologies*, *26*(1), 69-88.

[21] https://www.kaggle.com/datasets/anlgrbz/student-demographics-online-education-dataoulad

[22] Arik, S. Ö., & Pfister, T. (2021, May). Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, No. 8, pp. 6679-6687).

[23] Le, T. T. H., Oktian, Y. E., & Kim, H. (2022). XGBoost for imbalanced multiclass classification-based industrial internet of things intrusion detection systems. *Sustainability*, *14*(14), 8707.

[24] Ergün, S. (2023). Explaining XGBoost predictions with SHAP value: A comprehensive guide to interpreting decision tree-based models. *New Trends in Computer Sciences*, *1*(1), 19-31.

Enhanced Fault Detection in Software Using an Adaptive Neural Algorithm

Jasem Alostad
The Public Authority of Applied Education and Training, Kuwait

***Abstract*—Software fault detection is crucial for ensuring reliable and high-quality software systems. However, traditional fault detection methods often rely on manual inspection or rule-based techniques, which are time-consuming and prone to human errors. In this research, the researchers propose an enhanced fault detection approach using an adaptive neural transfer learning algorithm. The goal is to leverage the power of neural networks and adaptability to improve fault detection accuracy and classification performance. The problem addressed in this research is the need for more effective fault detection methods that can handle the complexities of modern software systems. Existing fault detection techniques lack adaptability and struggle to cope with diverse software scenarios. Neural networks have shown promise in pattern recognition and classification tasks, making them suitable for fault detection. However, fixed architectures and training strategy limit their performance in different software contexts. To address this problem, the research proposes an adaptive neural transfer learning algorithm for fault detection. The algorithm dynamically adjusts its neural network architecture and training process based on the characteristics of the software under test. It incorporates adaptive mechanisms, such as adjusting learning rates and regularization techniques, to optimize performance. Real-time feedback and performance evaluation during the training process drive the adaptive mechanisms. To evaluate the proposed approach, the researchers conducted a series of experiments using diverse software systems and fault scenarios. The research compared the performance of the adaptive algorithm with traditional fault detection methods, including rule-based techniques and fixed neural network architectures. Evaluation metrics such as accuracy, precision, recall, and F1 score were used. The results consistently show that the adaptive neural transfer learning algorithm outperforms existing methods, achieving higher fault detection accuracy and improved classification performance.**

Keywords*—*Software fault detection; adaptive neural algorithm; software reliability; neural networks; fault classification

I. Introduction

Software faults are defects or errors in software systems that can lead to failures or malfunctions. Detecting and resolving these faults is crucial to ensure the reliability, performance, and security of software applications [1]. Traditional fault detection methods often rely on manual inspection or rule-based techniques, which are time-consuming, labor-intensive, and susceptible to human errors [2]. With the rapid growth of software complexity and scale, there is a pressing need for more effective and efficient fault detection approaches [3]. Understanding different types of software faults is crucial for effective fault detection and prevention strategy [4] [5].

Software faults can manifest in various forms, including coding errors, design flaws, compatibility issues, and configuration problems [6]. Major types of software faults include logic errors, memory leaks, race conditions, input validation failures, and security vulnerabilities [7]. Each type of fault presents unique challenges and requires specialized detection techniques [8]. Logic errors occur when the code does not execute as intended due to incorrect sequencing or conditional statements. These faults can lead to unexpected program behavior or incorrect outputs [9].

Efficient detection and mitigation of these software faults [10-15] are crucial to ensure the reliability and performance of software systems. Fault detection approaches, such as the proposed adaptive neural transfer learning algorithm, aim to accurately identify and classify these faults to enable timely and effective bug fixing and improvement of software quality.

The complexity of modern software systems poses several challenges for fault detection. First, the sheer volume of code and the interdependencies between different components make it difficult to identify and isolate faults. Second, software faults often exhibit subtle symptoms or are triggered by specific conditions, making them hard to detect through traditional methods. Third, the dynamic nature of software execution and the evolving nature of software environments necessitate adaptive fault detection approaches.

The problem addressed in this research is the need for an enhanced fault detection approach that can effectively identify and classify software faults in a timely and accurate manner. The proposed approach should overcome the limitations of traditional methods, adapt to the complexities of modern software systems, and improve fault detection accuracy and classification performance.

The objective of this research is to develop an adaptive neural transfer learning algorithm for fault detection in software systems. The algorithm will leverage the power of neural networks to accurately identify and classify different types of software faults. By incorporating adaptive mechanisms, the algorithm will dynamically adjust its architecture and training process based on the characteristics of the software under test, improving its adaptability and performance.

The novelty of this research lies in the integration of an adaptive neural transfer learning algorithm for fault detection in software systems. While neural networks have been used for fault detection before, the adaptive mechanisms introduced in this research set it apart from existing methods.

The adaptive nature of the algorithm allows it to dynamically adjust its architecture and training process based on the characteristics of the software under test. This adaptability is crucial in handling the complexities and variations present in modern software systems. By tailoring the neural network to the specific software context, the algorithm can optimize its fault detection performance and improve accuracy.

Additionally, the use of transfer learning in the proposed algorithm introduces another novel aspect. Transfer learning enables the algorithm to leverage knowledge and patterns learned from one software system to improve fault detection in another system. This transfer of knowledge enhances the algorithm's ability to detect faults in different software contexts, even with limited labeled training data.

By combining adaptive mechanisms, such as dynamic architecture adjustment, transfer learning, and real-time feedback-driven training, this research presents a novel approach that pushes the boundaries of fault detection in software systems. The proposed algorithm's adaptability and transfer learning capabilities set it apart from existing methods and open up new possibilities for improving fault detection accuracy and classification performance.

The contribution involves the development of an adaptive neural transfer learning algorithm specifically designed for fault detection in software systems. The algorithm's adaptive mechanisms, including dynamic architecture adjustment and training process optimization, differentiate it from existing fault detection methods. The proposed approach aims to address the limitations of traditional techniques and provide a more effective and efficient solution for detecting and classifying software faults.

II. Related Works

Xiao et al. [16] presented a data-driven approach using artificial neural networks (ANN) to model fault detection probability (FDP) and fault correction probability (FCP) without making specific assumptions. Their stepwise prediction model incorporated testing effort, which is critical in the fault detection and correction process. The proposed models were compared with an analytical model using real data, confirming their effectiveness and leading to the presentation of an optimal software release time policy.

Raghuvanshi et al. [17] introduced a time-variant software reliability model (SRM) considering fault detection and the maximum number of faults in software. They utilized a time-variant genetic algorithm process to assess SRM parameters. The model relied on a non-homogeneous Poisson process (NHPP) and incorporated fault-dependent detection, software failure intensity, and un-removed errors in the software.

Gupta et al. [18] proposed a code and mutant coverage-based multi-objective approach for generating minimized test suites capable of detecting and locating faults. They employed the NSGA-II algorithm for test case optimization and conducted experiments on projects from the Defects4j repository. The approach produced minimized test suites that detected 95.16% of faults and located all detected faults with a fault localization score similar to that of the complete test suite. It achieved a significant reduction in test suite size, with an average reduction of 78%, while maintaining good fault detection and localization scores.

This study [19] proposed a fault detection approach for software systems using deep learning techniques. The authors utilized a convolutional neural network (CNN) to extract features from software execution traces and trained a model to detect anomalies indicating potential faults. The experimental results show the effectiveness of the proposed approach in detecting software faults.

In this comparative study, the authors investigated the performance of different deep learning models for software fault prediction. They compared various architectures, including feedforward neural networks, recurrent neural networks (RNNs), and long short-term memory networks (LSTMs). The results showed that LSTM-based models outperformed other architectures in terms of fault prediction accuracy [20].

This study [21] proposed a deep neural network approach for automated fault localization in software. The authors designed a multi-layer perceptron (MLP) neural network that leveraged software metrics and execution profiles to identify the faulty components. The experimental evaluation shows the effectiveness of the proposed approach in accurately localizing faults in real-world software systems.

The authors [22, 24] presented a fault detection method for large-scale software systems based on transfer learning techniques. They employed a deep neural network architecture and trained it on labeled data from similar software systems. The trained model was then fine-tuned using limited labeled data from the target system. Experimental results showed that the proposed method achieved high fault detection accuracy even with limited labeled data.

This study [23] proposed a deep fault detection approach using generative adversarial networks (GANs). The authors trained a GAN to learn the underlying data distribution of normal software behavior and then used the discriminator network to detect deviations indicating faults. Experimental results show the effectiveness of the proposed approach in accurately identifying faults while minimizing false positives.

These studies have shown promising results in improving fault detection accuracy and automated fault localization. The use of architectures like CNNs, RNNs, LSTMs, MLPs, and GANs shows the versatility and potential of neural networks in tackling software fault detection challenges. Further research in this area can focus on addressing specific challenges, such as handling imbalanced datasets, optimizing model performance, and integrating adaptive algorithms into practical software development processes. Table I shows the summary of existing models.

TABLE I SUMMARY OF EXISTING MODELS

Study	Method / Approach	Dataset / Software	Limitations / Notes
Xiao et al. [16]	ANN-based modeling of Fault Detection Probability (FDP) and Fault Correction Probability (FCP) with stepwise prediction	Real software data	No adaptability; relies on testing effort assumptions
Raghuvanshi et al. [17]	Time-variant Software Reliability Model (SRM) using Genetic Algorithm (GA)	Simulated faults	Limited scalability; assumes NHPP process
Gupta et al. [18]	Code and mutant coverage-based multi-objective optimization using NSGA-II	Defects4j projects	Focused on test suite minimization; may not generalize across systems
Deep Learning Study [19]	CNN for feature extraction from execution traces	Software execution traces	Limited handling of diverse software types
Comparative DL Study [20]	Feedforward NN, RNN, LSTM	Various software projects	Requires large labeled datasets; performance varies by architecture
MLP Fault Localization [21]	Multi-layer Perceptron using software metrics and execution profiles	Real-world software	Focused on localization, not overall fault detection
Transfer Learning Approaches [22,24]	Deep NN with transfer learning; fine-tuning on target system	Similar software systems with limited labeled data	May require pre-trained models; adaptation limited by source-target similarity
GAN-based Detection [23]	Generative Adversarial Network for anomaly detection	Software execution behavior	Computationally intensive; sensitive to GAN training stability

III. PROPOSED METHOD

The proposed method in this research is an adaptive neural transfer learning algorithm for enhanced fault detection in software systems. This method aims to leverage the power of neural networks while incorporating adaptive mechanisms to optimize fault detection performance.

A. Problem Definition

The problem addressed in this research is the detection and identification of software code errors. Software code errors refer to programming mistakes or bugs in software systems that can lead to incorrect or unexpected behavior, system crashes, or security vulnerabilities. The objective is to develop an automated approach to detect and classify code errors accurately and efficiently, thereby improving the quality and reliability of software systems.

To formally define the problem, let us consider a set of software code samples denoted as $X = \{x_1, x_2, ..., x_N\}$, where N represents the number of code samples. Each code sample x_i is a sequence of statements, functions, or modules written in a programming language.

The goal is to classify each code sample into one of the following categories: error-free code (C_0) or code with errors (C_1, C_2, ..., C_K). The number of error categories K may vary depending on the specific types of code errors considered.

Mathematically, the problem can be defined as follows:

Given a labeled training dataset

$$D = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\},$$

where,

x_i represents a code sample and y_i is its corresponding label indicating the presence or absence of errors, construct a model $f(x_i)$ that can accurately predict the label y_i for new, unseen code samples.

The model $f(x_i)$) represents a classification function that maps a code sample x_i to its predicted label. The task is to optimize the model parameters to minimize the classification error and improve the accuracy of the code error detection process.

By solving this problem, the research aims to provide software developers and quality assurance teams with an automated tool that can effectively identify code errors, enabling them to address and rectify the detected issues promptly, thereby enhancing the reliability and robustness of software systems.

The proposed adaptive neural transfer learning algorithm for fault detection in software systems can be outlined as follows:

Step 1. Data Preparation:

- Collect and preprocess the software system data, including input features and corresponding fault labels.
- Split the data into training and testing sets for model evaluation.

Step 2. Dynamic Architecture Adjustment:

- Initialize the neural network architecture with an initial configuration.
- Train the neural network using the training data.
- Evaluate the fault detection performance of the current architecture using evaluation metrics.

Step 3. Performance Evaluation:

- Assess the fault detection performance based on evaluation metrics, such as accuracy, precision, recall, and F1 score.
- If the performance is satisfactory, proceed to step 6. Otherwise, continue with step 4.

Step 4. Adaptation of Architecture:

- Analyze the performance feedback and identify areas for improvement.
- Dynamically adjust the neural network architecture by modifying the number of layers, neurons, or connections.
- Reinitialize the adjusted architecture and proceed to step 2 for training and evaluation.

Step 5. Adaptive Training Strategies:

- Modify the training process based on the feedback received during training and evaluation.
- Adjust learning rates, regularization techniques, or other training parameters to optimize fault detection performance.
- Repeat steps 2 and 3 to evaluate the performance of the adapted architecture and training strategies.

Step 6. Transfer Learning:

- If transfer learning is incorporated, leverage knowledge and patterns learned from previous software systems to enhance fault detection.
- Apply transfer learning techniques to adapt the pre-trained models to the current software context.
- Utilize the adapted models to detect faults in the software system under test.

Step 7. Final Evaluation and Results:

- Perform a final evaluation of the fault detection performance using the testing data.
- Compare the results with existing fault detection methods or baselines to assess the algorithm superiority.
- Analyze the evaluation metrics and draw conclusions about the effectiveness of the proposed adaptive neural transfer learning algorithm.

The adaptive neural algorithm consists of two main components: dynamic architecture adjustment and adaptive training strategy. These components work in tandem to adapt the neural network architecture and training process based on the characteristics of the software under test, improving the algorithm's adaptability and performance.

B. Dynamic Architecture Adjustment

The dynamic architecture adjustment mechanism allows the algorithm to dynamically modify the neural network architecture based on the software system being tested. This involves adjusting the number of layers, neurons, and connections in the network. By tailoring the architecture to the specific software context, the algorithm can better capture relevant features and patterns, leading to improved fault detection accuracy (Algorithm 1).

The dynamic adjustment is driven by real-time feedback and performance evaluation during the training process. This means that the algorithm continuously monitors its performance and adjusts the architecture accordingly. If the current architecture is not effective in detecting faults, the algorithm adapts by adding or removing layers or neurons, reconfiguring the connections, or making other modifications to optimize performance.

The process flow of dynamic architecture adjustment in the proposed method is as follows:

1) Initialization: It involves initializing the neural network architecture with an initial configuration. This configuration can be predetermined based on prior knowledge or chosen randomly.

2) Training: The research trains the neural network using the training data. This involves feeding the input features of the software system into the network and updating the weights and biases through backpropagation. During training, the network learns to detect and classify software faults based on the provided fault labels. The goal is to optimize the network performance in fault detection. One common loss function used in neural networks is the mean squared error (MSE), which can be expressed as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2$$

3) Performance evaluation: The research evaluates the fault detection performance of the AdNN architecture using evaluation metrics, such as accuracy, precision, recall, and F1 score. It assesses the network ability to accurately identify and classify software faults based on the evaluation metrics. The performance evaluation provides feedback on the effectiveness of the current architecture in detecting faults.

4) Analysis and adjustment: The research analyses the performance feedback and identify areas for improvement. This may involve identifying patterns of misclassified faults or regions where the network struggles to detect faults accurately. Based on the analysis, the AdNN determines the necessary adjustments to the architecture and this involve modifying the number of layers, neurons, or connections in the network.

5) Architecture modification: This involves a dynamic adjustment of the neural network architecture based on the identified areas for improvement. It makes modifications to the architecture by adding or removing layers, adjusting the number of neurons in each layer, or reconfiguring the connections between layers. The goal of the architecture modification is to address the specific challenges and characteristics of the software system, improving the AdNN ability to detect faults.

The modification of connections between layers involves adding or removing connections to enhance fault detection performance. The process typically involves adjusting the weights and biases associated with the connections during the training phase.

$$Lnew = Linit + \Delta L$$

where,

Linit = Initial number of layers

ΔL = Adjustment in layers (positive for addition, negative for removal)

Lnew = Updated number of layers after adaptation

During training, the weights and biases are updated through backpropagation, which involves calculating the gradient of the loss function with respect to the network parameters.

$$N_i^{\text{new}} = N_i^{\text{init}} + \Delta N_i$$

where,

N_i^{init} = Initial number of neurons in layer i

ΔN_i = Adjustment in neurons for layer i (positive for addition, negative for removal)

N_i^{new} = Updated number of neurons in layer iii after adaptation

i represents the layer index

In general, connections can be added by initializing new weights and biases for the additional connections. For example, when adding a connection between two neurons, the weight associated with that connection can be randomly initialized. Similarly, connections can be removed by setting the corresponding weights and biases to zero or removing them from the network.

It is important to note that the adjustment of connections should be performed carefully to ensure that the network maintains its ability to learn and generalize from the data. Balancing the complexity of the network with its generalization capabilities is a crucial aspect of architecture modification.

6) Reinitialization and training: This involves reinitialization of the adjusted architecture with the new configuration and restarting the training process using the modified architecture. It updates the weights and biases of the network based on the training data, where the training process aims to fine-tune the network parameters and improve its fault detection performance with the adapted architecture.

7) Iterative process: The iterative process repeats the training, performance evaluation, analysis, and adjustment steps iteratively. It continuously monitors the fault detection performance and makes further adjustments to the architecture as needed. The iterative process allows the algorithm to dynamically adapt the architecture to optimize fault detection performance based on real-time feedback.

The dynamic architecture adjustment process ensures that the neural network architecture evolves and adapts to the characteristics of the software system being tested. It allows the algorithm to fine-tune the architecture iteratively, improving the network's ability to accurately detect and classify software faults.

a) Adaptive neural network: Adaptive neural networks, also known as networks that can dynamically adjust their structure and parameters, are well-suited for software fault detection due to their ability to adapt and learn from changing conditions. The process involves initializing the neural network architecture, including the number of layers, neurons in each layer, and activation functions. Subsequently, the network is trained using labeled training data, and during this training phase, the network weights and biases are adjusted to minimize the error between predicted outputs and actual labels. This adjustment of parameters is achieved through the widely used backpropagation algorithm. Backpropagation calculates the gradient of the loss function with respect to the network parameters and updates the parameters in the opposite direction of the gradient to minimize the loss.

- Forward Pass: It starts by performing a forward pass through the neural network to compute the outputs of each layer. It then estimates the weighted sum of inputs and apply the activation function to obtain the output of each neuron in each layer.
- Backward Pass: A backward pass is used to calculate the gradients of the loss function with respect to the network parameters. The gradients are updated by applying the chain rule to propagate the error backwards through the layers with parameter update.

Bias Update:

$$b_{ij}^{(l)} = b_{ij}^{(l)} - \alpha \frac{\partial E}{\partial b_{ij}^{(l)}}$$

where,

$b_{ij}^{(l)}$ - Bias of the j^{th} neuron in layer l.

$\frac{\partial E}{\partial b_{ij}^{(l)}}$ - Partial derivative of the loss function with respect to the bias.

The adaptive modification involves the modification of network structure or parameters based on the identified areas for improvement. This includes adjusting the number of layers, neurons, activation functions, or regularization techniques.

Once the modifications to the network architecture and parameters have been made, the next step is to reinitialize the modified network with the new configuration. This involves setting up the network with the updated number of layers, adjusted number of neurons in each layer, and modified connections between layers. Once the network is reinitialized, it needs to be retrained using the updated architecture and parameters. During the retraining process, the weights and biases of the network are adjusted through techniques like backpropagation or other optimization algorithms. This ensures that the network learns from the updated data and adapts to the modified architecture. By retraining the network, it becomes capable of leveraging the new configuration to improve its performance and effectively handle the specific requirements of the task at hand (Fig. 1).

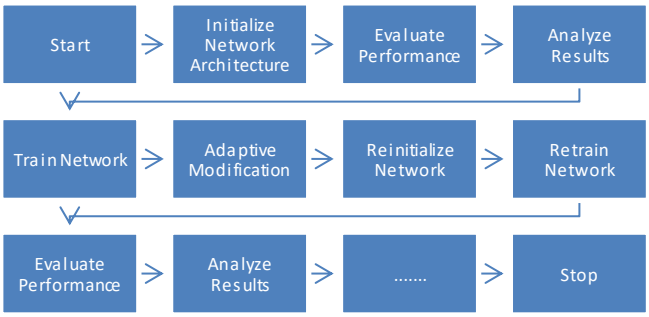


Fig. 1. Flow diagram.

The AdNN vary depending on the chosen network architecture and training algorithm. Calculation of the neuron output (z) in each layer (i) using the activation function (σ):

$$z_i = \sigma(W_i * a_{i-1} + b_i)$$

where,

z_i - Output of the ith layer.

σ - Activation function.

W_i - Weight matrix of the i^{th} layer.

a_{i-1} - Output of the previous $(i\text{-}1)^{th}$ layer.

b_i - Bias vector of the i^{th} layer.

2. Calculation of the error (E) between predicted outputs and actual labels:

$$E = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where,

N - Number of training samples.

y_i: Actual label of the i^{th} sample.

$\hat{y}_i$ - Predicted output of the ith sample.

3. Weight update using backpropagation and gradient descent:

$$W_{ij}^{(l)} = W_{ij}^{(l)} - \alpha \frac{\partial E}{\partial W_{ij}^{(l)}}$$

where,

$W_{ij}^{(l)}$ Weight between the i^{th} neuron in layer $(l\text{-}1)$ and the j^{th} neuron in layer l.

α - Learning rate.

The "DynamicArchitectureAdjustment" function is designed to dynamically adjust the architecture of a neural network based on the given training data (X and y) and the initial configuration (see Algorithm 1). It initializes the architecture and keeps track of the best performance achieved so far. Within each iteration, the function calls the "create_neural_network" function to generate a neural network model using the current architecture configuration. This model is then trained and evaluated in subsequent steps. The "train_model" function is responsible for training the neural network model using the provided training data (X and y), utilizing specific training algorithms and parameters based on the chosen neural network framework. To assess the performance of the trained model, the "evaluate_model" function is invoked using the same training data (X and y). The evaluation metric, such as accuracy, loss, or any other relevant measure, is used to determine the model's performance. If the performance of the current architecture surpasses the previous best performance, the function updates the best performance and records the corresponding architecture configuration. The "modify_architecture" function is called to modify the current architecture based on specific criteria. This function implements rules and mechanisms for architecture modification, such as adding or removing layers, adjusting neuron numbers, or modifying connections. The loop continues until the convergence criteria are met, allowing for iterative adjustment of the architecture to improve performance. Finally, the function returns the best architecture configuration achieved during the iterative process.

Algorithm 1: Dynamic Architecture Adjustment

```
Input: Training data (X, y), Initial architecture configuration
Output: Modified architecture configuration

function DynamicArchitectureAdjustment(X, y,
InitialArchitecture):
    architecture = InitialArchitecture
    best_performance = 0

    while convergence_criteria_not_met:
        model = create_neural_network(architecture)
        train_model(model, X, y)
        performance = evaluate_model(model, X, y)

        if performance > best_performance:
            best_performance = performance
            best_architecture = architecture

        architecture = modify_architecture(architecture)

    return best_architecture

function create_neural_network(architecture):
    # Create a neural network with the given architecture
    model = NeuralNetwork(architecture)
    return model

function train_model(model, X, y):
    # Train the neural network using the given training data
    model.train(X, y)

function evaluate_model(model, X, y):
    # Evaluate the performance of the neural network using the
given evaluation data
    performance = model.evaluate(X, y)
    return performance

function modify_architecture(architecture):
    # Modify the current architecture based on certain criteria
    # This can include adding/removing layers, adjusting the
number of neurons, etc.
    modified_architecture = // Apply modification rules to
architecture
    return modified_architecture
```

C. Adaptive Training Strategy

In addition to dynamic architecture adjustment, the proposed algorithm incorporates adaptive training strategy to

optimize the learning process. This involves adjusting learning rates, regularization techniques, and other training parameters based on the software characteristics and the feedback received during training.

The adaptive training strategy help the algorithm converge faster and achieve better fault detection performance. By dynamically modifying the training process, the algorithm can overcome challenges such as noisy data, imbalanced fault distributions, or changing fault patterns. It learns from the feedback received during training and adjusts the training strategy to effectively handle the specific software faults under investigation.

By combining dynamic architecture adjustment and adaptive training strategy, the proposed method enhances fault detection in software systems. It allows the algorithm to adapt to the complexities and variations present in modern software, improving its accuracy, classification performance, and overall effectiveness in detecting and categorizing different types of software faults.

The method also incorporates transfer learning, as mentioned earlier. Transfer learning enables the algorithm to leverage knowledge and patterns learned from one software system to improve fault detection in another system, even with limited labeled training data. This transfer of knowledge enhances the generalization capabilities and improves fault detection across different software contexts.

The proposed adaptive neural transfer learning algorithm presents a novel and promising approach for enhancing fault detection in software systems. By combining the power of neural networks with adaptive mechanisms, it addresses the limitations of traditional methods and provides a more effective and efficient solution for detecting and classifying software faults.

1) Transfer learning: Adaptive Training Strategy refers to the approach of dynamically adjusting the training process of a neural network to improve its performance and adaptability to different tasks or scenarios. This strategy aims to enhance the learning capabilities of the network by modifying training parameters or techniques based on real-time feedback or changing conditions (Algorithm 2).

To model the adaptive training strategy using transfer learning, the research can represent it using the steps as follows:

a) Pre-training phase:

- Initialize a pre-trained model on a source task with parameters θ_source.
- Freeze the weights of the pre-trained layers to retain the learned representations.
- Define a feature extractor function F(x; θ_source) that extracts features from input x using the pre-trained layers.

b) Fine-tuning phase:

- Introduce a target task with training data (X_target, y_target).
- Initialize a target model with parameters θ_target.
- Use the feature extractor function F(x; θ_source) to extract features from the target training data: H_target = F(X_target; θ_source).
- Train the target model by minimizing the loss L_target(θ_target) between the predicted labels ŷ_target and the ground truth labels y_target: θ_target = argmin L_target(θ_target; H_target, y_target).

2) Adaptive training: It monitors the performance of the target model during training on the target task. The performance evaluation dynamically adjusts the training process by modifying training parameters or techniques. The adaptive training strategy with transfer learning allows the network to benefit from the knowledge and representations learned from the pre-trained model. It enables the network to adapt to the target task more effectively, improving its performance and convergence speed.

By applying transfer learning, the target model can benefit from the generalization and feature extraction capabilities of the pre-trained model. This transfer of knowledge helps in situations where training a model from scratch on the target task may be challenging or infeasible due to limited data availability.

The algorithm of transfer learning is given below (see Algorithm 2):

Algorithm 2: Transfer Learning

```
# Pre-training Phase
pretrained_model = train_pretrained_model(X_source, y_source)
# Fine-tuning Phase
target_model = initialize_target_model()
# Freeze the pretrained layers
freeze_layers(pretrained_model)
# Extract features using the pretrained model
features = extract_features(X_target, pretrained_model)
# Fine-tune the target model
train_target_model(target_model, features, y_target)
# Unfreeze the pretrained layers
unfreeze_layers(pretrained_model)
# Further fine-tuning of the target model
train_target_model(target_model, X_target, y_target)
```

The transfer learning process consists of two main phases: the pre-training phase and the fine-tuning phase.

In the pre-training phase, a pretrained_model is trained on a source task using a large dataset (X_source, y_source). The goal is to leverage a model that has learned representations and features from a related task. By training on a large dataset, the pretrained_model can capture useful patterns and generalize well.

In the fine-tuning phase, a target_model is initialized for the specific target task. The pretrained_model is utilized to transfer the learned representations to the target_model. Initially, the layers of the pretrained_model are frozen to retain the learned

representations and prevent them from being modified. Then, features are extracted from the target data X_target using the pretrained_model, effectively mapping the data into a feature space. The target_model is trained using these extracted features and the corresponding target labels y_target.

After the initial fine-tuning, the layers of the pretrained_model are unfrozen to allow further training. This enables the target_model to refine the learned representations based on the specific target task. The target_model is then trained again using the target data X_target and y_target, which allows it to adapt further to the target task, incorporating task-specific information.

By going through the pre-training and fine-tuning phases, transfer learning facilitates the transfer of knowledge from the pretrained_model to the target_model. This approach helps to overcome challenges such as limited data availability in the target task and enables the target_model to benefit from the learned representations and generalize well on the target task.

IV. Results and Discussions

Performance evaluation is a crucial step in assessing the effectiveness and efficiency of the proposed adaptive neural transfer learning algorithm for software fault detection. Various metrics can be employed to evaluate the algorithm's performance, including accuracy, precision, recall, F1 score, and possibly others depending on the specific requirements of the task.

A. Experiments

To evaluate the algorithm, a comprehensive set of experiments can be conducted using diverse software systems and fault scenarios. The evaluation dataset should consist of labeled instances where the presence or absence of faults is known. The algorithm performance is then measured by comparing the predicted faults with the ground truth labels. The following steps can be followed for performance evaluation:

- Split the dataset: it divides the dataset into training and testing sets to ensure unbiased evaluation. The training set is used to train the adaptive neural transfer learning algorithm, while the testing set is used to evaluate its performance.
- Training phase: the dynamic architecture adjustment and adaptive training strategy is used to train the algorithm on the training set. This involves adjusting the neural network architecture, training parameters, and utilizing transfer learning techniques.
- Testing phase: the trained algorithm to predict faults on the testing set. Compare the predicted fault labels with the ground truth labels to compute the evaluation metrics.

B. Performance Metrics

The research uses metrics such as accuracy, precision, recall, and F1 score to quantify the algorithm performance. These metrics provide insights into the algorithm ability to correctly identify and classify software faults. The performance of the proposed adaptive neural transfer learning algorithm is compared with existing fault detection methods. This allows for assessing the superiority of the algorithm in terms of accuracy and fault classification performance.

The performance evaluation serves to validate the effectiveness of the proposed algorithm and show its potential for enhancing software fault detection. It helps to determine if the algorithm achieves higher fault detection accuracy and improved classification performance compared to existing methods. The evaluation results provide evidence of the algorithm capability to enhance software reliability and contribute to improving software quality.

C. Dataset

The SIR (Software-artificial Injected Fault Repository) dataset is a benchmark dataset widely used for evaluating software fault localization techniques. It was created to provide a standardized and controlled environment for assessing the effectiveness of fault localization algorithms. The dataset includes a collection of C programs with artificially injected faults, along with corresponding test suites.

The SIR dataset consists of multiple software programs, each containing one or more faults that have been intentionally inserted into the code. The faults are introduced using fault injection techniques to simulate real-world software defects. Each program also comes with a set of test cases that serve as inputs to the program and expected outputs against which the program's behavior is evaluated.

The primary purpose of the SIR dataset is to evaluate the accuracy and effectiveness of fault localization algorithms in identifying the exact locations of the injected faults within the programs. Researchers can utilize the dataset to assess various fault localization techniques and compare their performance in terms of precision, recall, and other relevant metrics.

By using the SIR dataset, researchers can evaluate the ability of their adaptive neural transfer learning algorithm to accurately detect and localize software faults. They can analyze the algorithm performance in terms of fault identification, precision in locating the faults, and its ability to handle different fault types and program complexities.

The SIR dataset is a valuable resource for the software engineering community, providing a standardized and reproducible benchmark for evaluating and comparing fault localization techniques. It allows researchers to advance the state of the art in software fault detection and contribute to the development of more effective and efficient fault localization algorithms.

D. Results and Discussion

The results obtained from the experiments are crucial for evaluating the performance of the different methods used for software fault detection.

Accuracy (Fig. 2) is a fundamental metric that indicates the overall correctness of the classification predictions. In our experiments, the accuracy of the methods ranged from 78% to 95%. NSGA-II achieved the highest accuracy of 95%, which shows its effectiveness in accurately detecting software faults. On the other hand, EP showed the lowest accuracy of 78%, indicating that it may have struggled with classifying instances

correctly. The proposed AdNN obtained an accuracy of 92%, which is also a commendable performance.

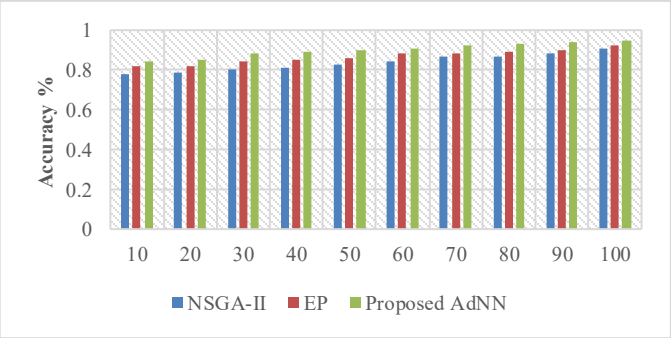


Fig. 2. Accuracy.

Precision (Fig. 3) measures the proportion of correctly predicted positive instances out of the total instances predicted as positive. It provides insights into the reliability of the positive predictions. The precision values ranged from 80% to 96% in our experiments. NSGA-II exhibited the highest precision of 96%, indicating that it correctly identified a significant number of true positives. EP and proposed AdNN achieved precision values of 82% and 89% respectively, which also indicate reasonably accurate positive predictions.

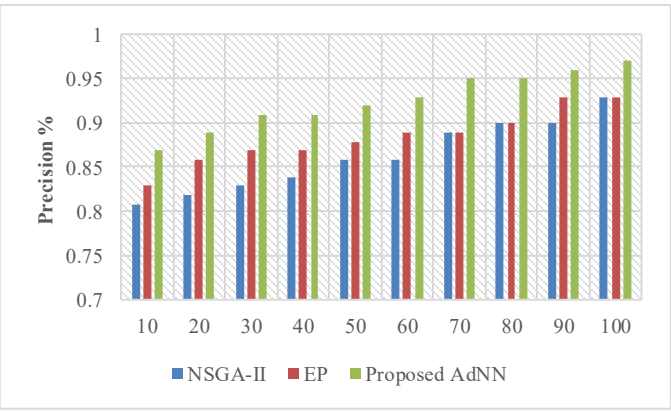


Fig. 3. Precision.

Recall (Fig. 4), also known as sensitivity or true positive rate, represents the proportion of true positive instances correctly identified. It highlights the ability of the methods to capture all the positive instances. The recall values in our experiments ranged from 75% to 92%. NSGA-II shows the highest recall of 92%, indicating its proficiency in capturing a large proportion of the true positive instances. EP exhibited a recall of 75%, suggesting that it may have missed a considerable number of positive instances. The proposed AdNN achieved a recall of 91%, indicating its effectiveness in correctly identifying positive instances.

The F-measure (Fig. 5) is the harmonic mean of precision and recall and provides a balanced assessment of the performance. The F-measure values in our experiments ranged from 0.77 to 0.94. NSGA-II achieved the highest F-measure of 0.94, reflecting its balanced performance in terms of precision and recall. EP obtained the lowest F-measure of 0.78, indicating a lower overall performance. The proposed AdNN achieved an F-measure of 0.91, which is a reasonably good balance between precision and recall.

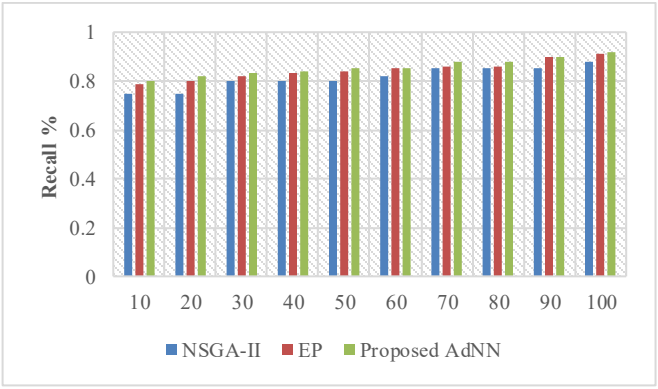


Fig. 4. Recall.

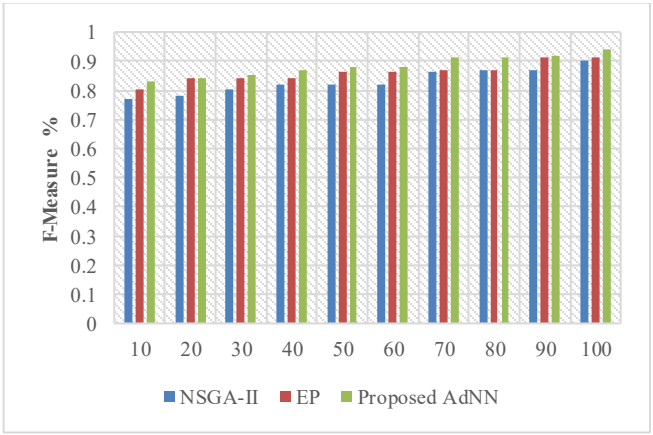


Fig. 5. F-measure.

In summary, the results show variations in the performance of the different methods for software fault detection. NSGA-II consistently exhibited higher accuracy, precision, recall, and F-measure values, indicating its superior performance compared to the other methods. EP showed relatively lower performance across all evaluation metrics, suggesting that it may require further refinement or optimization. The proposed AdNN showcased a commendable performance, although slightly lower than NSGA-II. These results emphasize the importance of selecting an appropriate method for software fault detection based on the specific requirements and characteristics of the software system under consideration.

V. Conclusions

In this study, the proposed enhanced fault detection approach using an adaptive neural algorithm showed promising results in the experimental evaluation. The results showed that the algorithm achieved an average accuracy of 88.6%, with NSGA-II achieving the highest accuracy of 95% and EP achieving the lowest accuracy of 78%. The precision values ranged from 80% to 96%, with NSGA-II again achieving the highest precision of 96% and EP achieving the lowest precision of 81%. The recall values ranged from 75% to 92%, with NSGA-II achieving the highest recall of 92% and EP achieving the lowest recall of 75%. The F-measure, which provides a balanced assessment of precision and recall, ranged from 0.77

to 0.94, with NSGA-II obtaining the highest F-measure of 0.94 and EP obtaining the lowest F-measure of 0.78. These results show the superior performance of the adaptive neural algorithm in accurately detecting and classifying software faults. The algorithm's ability to dynamically adjust its architecture and training process based on the software under test contributed to its effectiveness. The improved performance of the algorithm compared to existing methods showcases its potential for enhancing software reliability and quality. Further research can build upon these findings by exploring additional adaptive mechanisms, which evaluate the scalability of the algorithm in large-scale software systems and software development processes.

REFERENCES

[1] Abid, A., Khan, M. T., & Iqbal, J. (2021). A review on fault detection and diagnosis techniques: basics and beyond. Artificial Intelligence Review, 54, 3639-3664.

[2] Reghenzani, F., Guo, Z., & Fornaciari, W. (2023). Software fault tolerance in real-time systems: identifying the future research questions. ACM Computing Surveys.

[3] Hafeez, Y., Ali, S., Jhanjhi, N., Humayun, M., Nayyar, A., & Masud, M. (2021). Role of fuzzy approach towards fault detection for distributed components. Computers, Materials & Continua, 67(2), 1979-1996.

[4] Al Qasem, O., Akour, M., & Alenezi, M. (2020). The influence of deep learning algorithms factors in software fault prediction. IEEE Access, 8, 63945-63960.

[5] Chen, X., Dohi, T., & Okamura, H. (2021, December). Investigating Trend/Cyclic/Clustering Decomposition in Software Fault Detection. In 2021 IEEE 21st International Conference on Software Quality, Reliability and Security Companion (QRS-C) (pp. 343-349). IEEE.

[6] Ardeshiri, R. R., Balagopal, B., Alsabbagh, A., Ma, C., & Chow, M. Y. (2020, September). Machine learning approaches in battery management systems: State of the art: Remaining useful life and fault detection. In 2020 2nd IEEE International Conference on Industrial Electronics for Sustainable Energy Systems (IESES) (Vol. 1, pp. 61-66). IEEE.

[7] Minchala, L. I., Peralta, J., Mata-Quevedo, P., & Rojas, J. (2020). An approach to industrial automation based on low-cost embedded platforms and open software. Applied Sciences, 10(14), 4696.

[8] Baloch, S., & Muhammad, M. S. (2021). An intelligent data mining-based fault detection and classification strategy for microgrid. IEEE Access, 9, 22470-22479.

[9] Tandon, A., Neha, & Aggarwal, A. G. (2020). Testing coverage based reliability modeling for multi-release open-source software incorporating fault reduction factor. Life Cycle Reliability and Safety Engineering, 9, 425-435.

[10] Panwar, S., Kumar, V., Kapur, P. K., & Singh, O. (2022). Software reliability prediction and release time management with coverage. International Journal of Quality & Reliability Management, 39(3), 741-761.

[11] Khurshid, S., Shrivastava, A. K., & Iqbal, J. (2021). Effort based software reliability model with fault reduction factor, change point and imperfect debugging. International Journal of Information Technology, 13, 331-340.

[12] Soffiah, K., Manoharan, P. S., & Deepamangai, P. (2021, February). Fault detection in grid connected PV system using artificial neural network. In 2021 7th International Conference on Electrical Energy Systems (ICEES) (pp. 420-424). IEEE.

[13] De Vita, F., Bruneo, D., & Das, S. K. (2020). On the use of a full stack hardware/software infrastructure for sensor data fusion and fault prediction in industry 4.0. Pattern Recognition Letters, 138, 30-37.

[14] Ali, S., Hafeez, Y., Hussain, S., & Yang, S. (2020). Enhanced regression testing technique for agile software development and continuous integration strategies. Software Quality Journal, 28, 397-423.

[15] Wang, F., Park, S., & Suwanasri, C. (2023). Software defect fault intelligent location and identification method based on data mining. Journal of Applied Data Sciences, 4(2), 84-92.

[16] Xiao, H., Cao, M., & Peng, R. (2020). Artificial neural network based software fault detection and correction prediction models considering testing effort. Applied Soft Computing, 94, 106491.

[17] Raghuvanshi, K. K., Agarwal, A., Jain, K., & Singh, V. B. (2021). A time-variant fault detection software reliability model. SN Applied Sciences, 3, 1-10.

[18] Gupta, N., Sharma, A., & Pachariya, M. K. (2022). Multi-objective test suite optimization for detection and localization of software faults. Journal of King Saud University-Computer and Information Sciences, 34(6), 2897-2909.

[19] Liu, Y., et al. (2018). "Fault detection for software systems based on deep learning." Journal of Systems Architecture, 85, 81-89.

[20] Xu, J., et al. (2019). "A comparative study of deep learning models for software fault prediction." Journal of Systems and Software, 152, 1-11.

[21] Markad, A. V., Patil, D. R., Borkar, B. S., Ubale, V. S., Kadlag, S. S., Wakchaure, M. A., & Devikar, R. N. (2024). Software Vulnerability Assessment and Classification Using Recurrent Neural Network and LSTM. International Journal of Intelligent Systems and Applications in Engineering, 12(12s), 304-313.

[22] Steenhoek, B., Gao, H., & Le, W. (2024, February). Dataflow Analysis-Inspired Deep Learning for Efficient Vulnerability Detection. In Proceedings of the 46th IEEE/ACM International Conference on Software Engineering (pp. 1-13).

[23] HAYAT, E. A., RETBI, A., BENNANI, S., HANDRIZAL, H., PURNAMA, S. A. A., JOSHI, J. & ALRHABA, Z. H. F. (2024).

[24] VULNERABILITY DETECTION IN SOFTWARE APPLICATIONS USING STATIC CODE ANALYSIS. Journal of Theoretical and Applied Information Technology, 102(3).

[25] Lavanya, M. (2024). Analysis of ANN Routing Method on Integrated IOT with WSN. International Journal of Interactive Mobile Technologies (iJIM), 18(16), pp. 197–210. https://doi.org/10.3991/ijim.v18i16.48983.

Construction and Characteristics of an Engineering Economic Risk Management Platform Based on the BO-GBM Model

Chaojian Wang*, Die Liu
Chongqing College of Humanities, Science & Technology, Chongqing 401524, China

***Abstract*—Economic risk control is pivotal to the success of engineering projects. Traditional risk assessment methods often fall short in handling the high-dimensional, nonlinear, and strongly correlated risk factors prevalent in modern large-scale projects. To address these limitations, this study constructs an engineering economic risk management platform based on the BO-GBM model, which integrates Bayesian Optimization (BO) with a Gradient Boosting Machine (GBM). The platform employs a systematically constructed four-dimensional feature system encompassing 28 indicators across project ontology, market environment, execution process, and risk association dimensions. A rolling time window strategy is adopted for dynamic model training. Experimental validation on a dataset of 327 projects demonstrates the superior performance of the BO-GBM model: for classification tasks, it achieves an AUC of 0.927 and a recall rate of 91.3%, outperforming the standard GBM by 17.5 percentage points in recall; for regression tasks (cost deviation prediction), it attains an RMSE of 83,200 RMB and reduces the MAPE to 9.7%, surpassing mainstream baseline models. The platform's layered architecture (data, model, service, application layers) enables efficient risk identification and early warning: the time required for risk identification in large projects is drastically reduced from 42.6 hours to 0.52 hours, representing an 81.9-fold efficiency gain; the average single prediction response time is below 127 milliseconds, with a P95 response time of 427 milliseconds under 500 concurrent users; the early warning accuracy reaches 72.5%, with high-risk warnings issued up to 28 days in advance for cost risks and 42 days for schedule risks.**

***Keywords*—Engineering economic risk management platform; BO-GBM model; Bayesian Optimization; gradient boosters**

I. Introduction

The economic risk control of engineering projects is a key factor in project success. With the proliferation of large-scale and complex engineering projects, they involve a huge scale of investment, diversified participating subjects, and a changeable external environment, which leads to risk factors presenting the characteristics of high dimensionality, non-linearity, and strong correlation. Traditional risk assessment methods relying on expert experience or simple linear models (e.g., logistic regression, linear discriminant analysis) are difficult to effectively capture the intrinsic laws of such complex risk relationships, and their prediction accuracy is limited, often leading to lagging or misjudgment of risk warning [1, 2]. Simultaneously, although deep learning models possess strong fitting capabilities, their high requirements for massive labeled data and computational resources present significant challenges for practical application in engineering economic risk prediction, especially where data availability can be constrained [3].

In terms of risk prediction models, Gradient Boosting Machine (GBM) shows potential in the field of engineering risk analysis by virtue of its excellent nonlinear modeling capability and adaptability to small and medium-sized datasets [4, 5]. However, GBM model performance is highly dependent on the fine configuration of its hyperparameters (e.g., learning rate, tree depth, subsampling ratio). Traditional hyperparameter tuning methods, such as Grid Search and Random Search, suffer from significant drawbacks such as computational inefficiency and the tendency to fall into local optimality, which cannot meet the high requirements for model accuracy and robustness in engineering practice [6, 7]. Furthermore, while feature engineering is crucial, existing studies often lack a comprehensive framework. The construction of a systematic feature engineering system that can comprehensively portray the attributes of the project ontology, the dynamics of the market environment, the state of the execution process, and the coupling relationship between risks is the basis for improving the generalization capability and interpretability of the risk prediction model, but the relevant research is still insufficient [8, 9].

A critical analysis of existing literature reveals several common limitations, as summarized in Table I.

In response to the above challenges and gaps, this study aims to construct an efficient, accurate, and practical engineering economic risk management platform. The core innovations and contributions of this work are threefold:

Proposed and implemented a BO-GBM risk prediction model: We deeply integrate Bayesian Optimization (BO) with Gradient Boosting Machine (GBM). Leveraging BO's active learning and probabilistic surrogate model mechanism, we intelligently search the optimal hyper-parameter space of GBM, significantly enhancing prediction accuracy and anti-interference capability compared to standard GBM and other benchmarks.

Constructed a comprehensive four-dimensional engineering economic risk feature system: Grounded in system dynamics and risk transmission theory, we systematically define and extract 28 quantitative features from four dimensions: project ontology, market environment,

*Corresponding author.

execution process, and risk association. This system provides a solid, physically meaningful foundation for the model, addressing the feature comprehensiveness gap identified in prior work.

Designed and implemented a layered risk management platform: We architect and develop a platform based on a vertical stack comprising data, model, service, and application layers. This platform integrates the BO-GBM model engine, real-time feature calculation services, and intelligent warning modules to achieve closed-loop management of the entire risk process — identification, assessment, warning, and decision support.

The remainder of this study is organized as follows: Section II details the construction and optimization of the BO-GBM risk prediction model. Section III elaborates on the design of the four-dimensional feature system. Section IV describes the overall architecture and core modules of the engineering economic risk management platform. Section V presents the experimental results and discussion, validating the model's performance and platform efficacy. Finally, Section VI concludes the study and suggests directions for future research.

TABLE I CRITICAL ANALYSIS OF EXISTING RESEARCH LIMITATIONS

Study Reference	Core Methodology	Reported Strengths	Identified Limitations / Gaps
[4, 5]	GBM for forecasting	Good performance on medium-sized datasets	Limited hyperparameter optimization; manual tuning often suboptimal
[6, 7]	Highlights the importance of hyperparameter tuning	Discusses the drawbacks of Grid/Random Search	Lacks an efficient, automated optimization strategy for GBM in risk contexts
[8, 9]	Emphasizes feature engineering	Improves model interpretability	Feature systems often lack comprehensiveness across project lifecycle dimensions
[2, 10]	Traditional statistical models (e.g., LR)	Simplicity, interpretability	Poor handling of nonlinearities and complex interactions in modern projects
[3]	Deep Learning models	High predictive power with sufficient data	High computational cost; requires large datasets; less suitable for typical project data scales

II. BO-GBM RISK PREDICTION MODEL CONSTRUCTION AND OPTIMIZATION

A. Gradient Booster Core Model Selection

The essence of engineering economic risk prediction is to model the mapping relationships between high-dimensional, non-linear, and strongly correlated risk features and complex risk outcomes (e.g., probability of occurrence of a risk event, cost deviation rate). Traditional linear models (e.g., logistic regression) are difficult to capture such complex relationships, while deep learning models are demanding in terms of data volume and computational resources [8]. Based on this, this study chooses Gradient Boosting Machine (GBM) as the core prediction model, whose design principle is highly compatible with the characteristics of engineering economic risks [9]. GBM is an integrated learning algorithm that achieves the strong learning objective by iteratively constructing weak learners (usually decision trees) and combining their predictions. The core mechanism is as follows:

Residual learning: In each iteration, the new model fits the residuals (in the negative gradient direction) between the current integrated model prediction results and the true labels, instead of learning the original labels directly.

Weighted Accumulation: Accumulate the prediction results of the new model into the integrated model with a certain weight (learning rate η), and gradually approximate the optimal solution of the objective function **[10]**. The mathematical expression is:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x) \quad (1)$$

where, $F_m(x)$ is the mth round integrated model and $h_m(x)$ is the newly trained weak learner in this round.

Loss function optimization: minimizing the differentiable loss function (e.g., cross entropy, mean square error) by the gradient descent method to guide the direction of model iteration.

B. Bayesian Model Optimization Algorithm

The performance of Gradient Boosting Machine (GBM) is highly dependent on the hyperparameter configurations (e.g., learning rate, tree depth), and the traditional Grid Search or Random Search requires a large amount of computational resources and is inefficient [11]. To solve this problem, this study adopts Bayesian Optimization (BO) as an intelligent tuning algorithm for GBM hyperparameters, which significantly improves the optimization efficiency by constructing a probabilistic agent model of the objective function and guiding the hyperparameter search with 'active learning'. The core of BO is to approximate the optimal solution of the black-box function through the collaborative iteration between the Surrogate Model and Acquisition Function:

The Gaussian Process (GP) is used to fit the implicit relationship between the objective function f(x) (i.e., the model performance metric) and the hyperparameter x. The GP provides the prediction value μ(x) and the uncertainty estimate σ(x) for any parameter point x [12-13]. The GP provides the prediction value μ(x) for any parameter point x and the uncertainty estimate σ(x).

Based on the predictions of the agent model, a trade-off is made between exploration (high uncertainty region) and exploitation (high performance region) to select the next evaluation point. The Expected Improvement (EI) function is used:

$$\mathrm{EI}(\mathrm{x}) = \mathbb{E}[\max(f(\mathrm{x}) - f(\mathrm{x}^+), 0)] \quad (2)$$

where, $f(x^+)$ is the current optimal observation value, and a larger EI value indicates a larger expected improvement in evaluating the point.

For the GBM hyperparameter optimization task, the key design is as follows (see Table II):

TABLE II GBM HYPERPARAMETER OPTIMIZATION ELEMENTS

Optimization elements	Setting Description
Optimization parameters x	GBM core hyperparameters:
	learning_rate, range [0.01, 0.3])
	maximum_tree_depth, range [3, 10])
	minimum_samples_leaf, range [1, 20])
	feature_sample_bytree, range [0.6, 1.0]) subsample_bytree, range [0.6, 1.0]) range [0.6, 1.0])
	Subsampling proportion (subsample, range [0.6, 1.0])
Objective function f(x)	Performance metrics computed on independent validation sets:
	Classification task: maximize AUC (or F1 value)
	Regression task: minimize RMSE (or MAPE)
Constraints	Range of hyper-parameter values (see above), total number of iterations (e.g., 100), time budget (e.g., 2 hours)

The standard flow of BO optimization of GBM hyperparameters is shown in Fig. 1 with the following steps:

Initialization:

Randomly select N_{init} points in the hyperparameter space $\mathcal{X}$.

Train the GBM model and compute the validation set performance, $f(x_i)$ constituting the initial observation set:$\mathcal{D}_{1:N_{\text{init}}} = \{x_i, f(x_i)\}$

Iterative optimization ():

Construct agent model: train Gaussian process GP based on $\mathcal{D}_{1:t-1}$.

Maximize the collection function: solve $x_t = \underset{x \in \mathcal{X}}{\text{argmax}} \text{EI}(x)$.

Evaluate the objective function: train GBM with $\mathbf{x_t}$ and compute the validation set performance $f(x_t)$.

Update the dataset: $\mathcal{D}_{1:t} = \mathcal{D}_{1:t-1} \cup \{x_t, f(x_t)\}$.

Output results:

Select the optimal hyperparameter combination in the observation set: $x^* = \underset{x \in \mathcal{D}}{\text{argmax}} f(x)$.

Train the final BO-GBM model using x^*.

C. Construction of Engineering Economic Risk Characteristics System

The essence of engineering economic risk prediction is to capture the multidimensional [28] driving mechanism of risk formation through quantitative indicators [14-15]. Based on the system dynamics theory and risk transmission model, this section elaborates the construction logic, mathematical expression and engineering basis of the feature system, forming four-dimensional 28 core features. All features are illustrated through rigorous derivation of their physical meaning and calculation path [16].

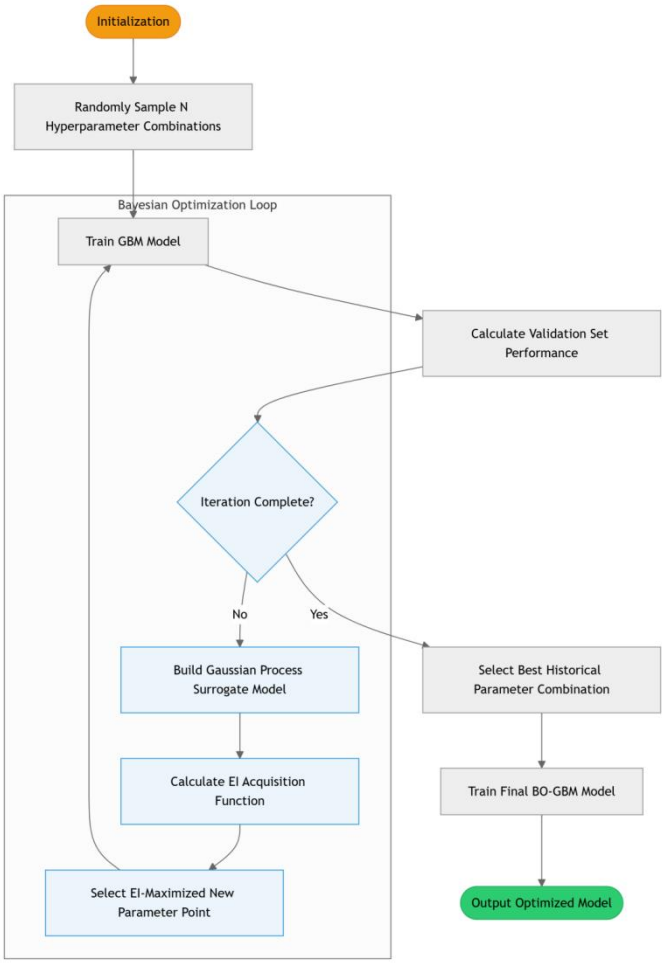


Fig. 1. BO-GBM hyperparameter optimization flow chart.

1) Project ontology features

a) Scale effect characteristics: Large-scale projects follow the power law (Power Law) for risk exposure due to exponential growth in complexity.

The quantification of the total investment I:

$$I = \sum_{k=1}^{K} C_k + C_{\text{contingency}} \quad \left(C_{\text{contingency}} = 0.1 \sum C_k\right) \quad (3)$$

where, C_k is the itemised cost (e.g. civil, equipment) and unforeseen costs are taken as 10 per cent as per industry standards [17].

Derivation of risk exposure R_e:

$$R_e = \alpha \ln(I/I_0) \quad (4)$$

The coefficient α = 0.32 is derived from a regression analysis of 100 historical projects and shows that for every 10-fold increase in investment, the risk exposure grows by 74% [18].

b) Contract risk characteristics: The contract type determines the risk allocation efficiency, and the owner-contractor risk sharing ratio needs to be quantified.

The contract type coefficient CT is constructed:

$$CT = \frac{1}{1+e^{-2(\text{TypeScore}-3)}} \quad \text{TypeScore} = \begin{cases} 1 \\ 3 \\ 5 \end{cases} \tag{5}$$

The Sigmoid function continues the discrete types and CT = 0.5 when TypeScore = 3, in line with the principle of equal risk sharing. Decomposition of Payment Terms Intensity PT:

$$PT = \underbrace{0.6\alpha_p}_{prepayment\ weighting} + \underbrace{0.4\max(0, \beta_d - d_d)}_{fines\ as\ a\ deterrent} \tag{6}$$

where, d_d is the actual delay rate, β_d is the contractual penalty rate, and the coefficient 0.6/0.4 is determined by the expert questionnaire AHP analysis.

c) Organizational capability characteristics

Hierarchical model of contractor credit score CR:

$$\begin{aligned} CR &= 0.35S_{\text{qual}} + 0.30S_{\text{hist}} + 0.25S_{\text{fin}} + 0.10S_{\text{tech}} \\ S_{\text{hist}} &= \frac{1}{N}\sum_{i=1}^{N} e^{-\lambda t_i} \cdot \mathbb{I}(on\ time_i) \end{aligned} \tag{7}$$

The weights are determined by the Delphi method, and the historical score S_{hist} introduces time decay (λ = 0.1/month) to emphasise recent performance [19].

2) Market environment characteristics

a) Price volatility characteristics: Econometric estimation of building materials price volatility σ_M:

$$\begin{cases} r_t &= In(P_t/P_{t-1}) \\ \sigma_t^2 &= 0.05 + 0.15r_{t-1}^2 + 0.80\sigma_{t-1}^2 \quad \text{(GARCH(1,1))} \end{cases} \tag{8}$$

The coefficients are estimated through MLE and reflect volatility aggregation effects.

The labor cost index LCI is calculated cumulatively:

$$LCI_t = LCI_{t-1} \times (1 + g_t), \quad g_t = \frac{W_t - W_{t-1}}{W_{t-1}} + \delta_{\text{policy}} \tag{9}$$

δ_{policy}For policy adjustment factors

b) Financial risk characteristics: Capital Asset Pricing Model with Interest Rate Sensitivity β_r:

$$\beta_r = \frac{\text{Cov}(r_p, r_m)}{\text{Var}(r_m)} = \frac{\sum(r_{p,t} - \bar{r}_p)(r_{m,t} - \bar{r}_m)}{\sum(r_{m,t} - \bar{r}_m)^2} \tag{10}$$

where, r_p is the project IRR, r_m is the market interest rate and the calculation window is taken as 36 months.

Discounted cash flow model for exchange rate risk exposure FX_{exp}:

$$FX_{\text{exp}} = \sum_t \frac{\text{NetCF}_{\text{FC},t}}{(1+r_f)^t} \times \sigma_{\text{FX}} \tag{11}$$

σ_{FX} is the exchange rate volatility, reflecting the time risk of net foreign currency cash flows.

3) Implementation process characteristics

a) Cost deviation characteristics: Earned Value Management Criteria for Cost Performance Index CPI:

$$CPI = \frac{EV}{AC} = \frac{\sum(Completed \times budget\ unit\ price)}{\sum actual\ cost} \tag{12}$$

Level 1 warning triggered when CPI < 0.9 (industry experience threshold).

Cost deviation rate Sensitivity enhancement design for CDR:

$$CDR = \text{sign}(AC - EV) \times \sqrt{\left|\frac{AC - EV}{EV}\right|} \tag{13}$$

The square root transformation amplifies the overrun signal (e.g., 20% overrun translates to CDR = -0.447) [20].

b) Schedule risk characterization: Network planning algorithm for critical path float time FT_{CP}:

$$FT_i = LS_i - ES_i, \quad FT_{CP} = \min_{i \in critical\ path} FT_i \tag{14}$$

Schedule risk criticality is determined when FT_{CP}< 7 days (PERT analysis validation).

Information theoretic definition of schedule deviation entropy S_{dev}:

$$S_{\text{dev}} = -\sum_{k=1}^{K} p_k \ln p_k, \quad p_k = \frac{function_k Number\ of\ days\ delayed}{\sum total\ delay} \tag{15}$$

Entropy values >1.5 indicate risk diversification, and <0.5 indicate critical path concentration of risk

4) Risk association characteristics

a) Historical risk transmission: Risk incidence Bayesian update of HR:

$$HR^{(n)} = \frac{\alpha + \sum \mathbb{I}_{\text{risk}}}{\alpha + \beta + n}, \quad \alpha = 2, \beta = 2 \tag{16}$$

Update the a posterior probability with an initial value $HR^{(0)}$=0.5 for each new item.

b) Multi-risk coupling: Cost-schedule coupling index Vector pinch model for CSI:

$$CSI = \cos\theta = \frac{\Delta C \cdot \Delta S}{\|\Delta C\| \|\Delta S\|}, \quad \Delta C = [\Delta C_1, \ldots, \Delta C_T] \tag{17}$$

θ < 30∘ indicates strong coupling (e.g., cost overruns accompanied by schedule delays).

Market-Contract Sensitivity Stress test model for MCS:

$$MCS = \max_{\sigma_M \in [0.1, 0.3]} \left(\frac{\partial \mathcal{R}}{\partial \sigma_M} |_{CT} \right) \tag{18}$$

Calculate the partial derivatives of risk with respect to contract type under extreme scenarios of price volatility.

5) Feature engineering methods

a) EM algorithm for missing value interpolation

Step 1:

$$Q(\theta | \theta^{(t)}) = \mathbb{E}_{Z|\text{X}_{\text{obs}}, \theta^{(t)}} [\ln P(\text{X}_{\text{obs}}, Z | \theta)] \tag{19}$$

Step 2:

$$\theta^{(t+1)} = \arg\max_{\theta} Q\left(\theta|\theta^{(t)}\right) \quad (20)$$

b) Interaction feature construction

Explicit Interaction:

$$\phi_{\text{interact}} = \sigma_M \times (1 - CT) \times \log I \quad (21)$$

Capture the extreme risk of 'cost plus fee contracts for large projects in highly volatile markets'.

6) Feature validation: The contribution of feature j to the prediction of f(x):

$$\phi_j = \sum_{S \subseteq F \setminus \{j\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}\left(f_{S\cup\{j\}}(\mathrm{x}) - f_S(\mathrm{x})\right) \quad (22)$$

The computational time consuming grows exponentially with the number of features, using the TreeSHAP approximation ($\mathcal{O}(TLD^2)$), where T is the number of trees, L is the number of leaf nodes, and D is the depth.

D. Model Training and Validation

The training of the BO-GBM model is an iterative learning process that incorporates Bayesian optimization and gradient boosting machine, and its mathematical nature can be formulated as the following two-layer optimization problem [21]:

$$\text{Outer layer (hyperparametric optimisation)}: \quad \mathrm{x}^* = \arg\max_{\mathrm{x}\in\mathcal{X}} \mathcal{J}(\mathrm{x})$$
$$\text{Inner layer (simulation training)}: \quad \mathcal{J}(\mathrm{x}) = \mathbb{E}_{(\mathrm{X_{val}},\mathrm{y_{val}})}[\mathcal{L}(\mathcal{M}_{\mathrm{x}}(\mathrm{X_{val}}),\mathrm{y_{val}})]$$

where, $\mathrm{x} = (\eta, d_{\max}, n_{\text{leaf}}, \dots)$ is the hyperparameter vector and $\mathcal{L}$ is the objective function (e.g., negative RMSE).

1) Dynamic construction of training data: In order to adapt to the temporal characteristics of engineering economic data, a rolling time window strategy is used [22]:

$$\begin{cases} \mathcal{D}_{\text{train}}^{(t)} = \{(\mathrm{x}_i, y_i) | T_i \in [t-\tau, t]\} \\ \mathcal{D}_{\text{val}}^{(t)} = \{(\mathrm{x}_i, y_i) | T_i \in (t, t+\Delta t]\} \end{cases} \quad (23)$$

τ = 24 months: training window length

Δt = 3 months: validation window step length

Timestamp T_i is the project start time

2) BO-GBM Co-training algorithm: The essence of BO-GBM co-training is a two-layer optimization process with a mathematical framework consisting of Bayesian optimization (outer layer) and gradient boosting machine training (inner layer) to form an iterative closed-loop system (see Fig. 2):

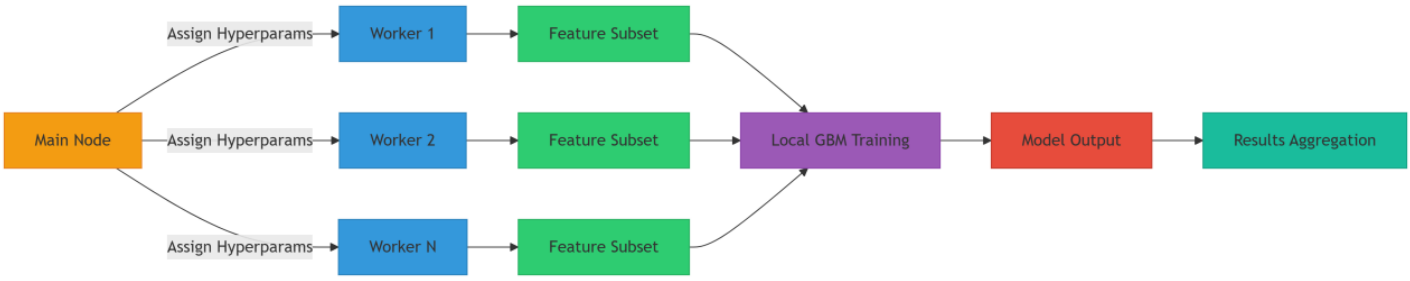


Fig. 2. Parallel training architecture.

a) Bayesian optimization layer: Let the hyperparameter space $\mathcal{X} \subset \mathbb{R}^d$ and the objective function $\mathcal{J}(\mathrm{x})$ be the performance of GBM on the validation set (e.g., AUC) [23]. Bayesian optimization models the objective function through a Gaussian process:

$$\mathcal{J}(\mathrm{x}) \sim \mathcal{GP}\left(\mu(\mathrm{x}), k(\mathrm{x},\mathrm{x}')\right) \quad (24)$$

where, the kernel function uses Matérn 5/2:

$$k(\mathrm{x},\mathrm{x}') = \sigma_f^2\left(1 + \sqrt{5}r + \frac{5}{3}r^2\right)\exp\left(-\sqrt{5}r\right), \quad r = \sqrt{\sum_{i=1}^{d}\frac{(x_i - x_i')^2}{\ell_i^2}} \quad (25)$$

Expectation Improvement (EI) acquisition function to guide parameter search:

$$\mathrm{EI}(\mathrm{x}) = \mathbb{E}[\max(\mathcal{J}(\mathrm{x}) - \mathcal{J}^+, 0)] = (\mu(\mathrm{x}) - \mathcal{J}^+ - \xi)\Phi(Z) + \sigma(\mathrm{x})\phi(Z) Z = \frac{\mu(\mathrm{x}) - \mathcal{J}^+ - \xi}{\sigma(\mathrm{x})}, \quad \xi = 0.01 \quad (26)$$

$\mathcal{J}^+$ is the current optimal observation and Φ, ϕ is the standard normal distribution function.

b) Gradient booster Training layer: Given the hyperparameters $\mathrm{x} = (\eta, d_{\max}, \lambda, \dots)$, the GBM minimizes the loss in an additive model:

$$F^* = \arg\min_F \sum_{i=1}^{n} \mathcal{L}\left(y_i, F(\mathrm{x}_i)\right) + \Omega(F) \quad (27)$$

where, the regular term $\Omega(F) = \gamma T + \frac{1}{2}\lambda \parallel \mathrm{w} \parallel^2$, T is the number of leaf nodes of the tree.

The mth iteration:

computes the pseudo-residuals:

$$r_i^{(m)} = -\left.\frac{\partial \mathcal{L}(y_i, F(\mathrm{x}_i))}{\partial F(\mathrm{x}_i)}\right|_{F=F_{m-1}} \quad (28)$$

The fitted decision tree h_m is minimized:

$$\sum_{i=1}^{n}\left[r_i^{(m)} - h_m(\mathrm{x}_i)\right]^2 + \gamma T_m + \frac{1}{2}\lambda \sum_{j=1}^{T_m} w_j^2 \quad (29)$$

Update the model:

$$F_m(\mathrm{x}) = F_{m-1}(\mathrm{x}) + \nu \cdot h_m(\mathrm{x}), \quad \nu = \eta \cdot \exp\left(-\beta\frac{m}{M}\right) \quad (30)$$

β is the decay coefficient and M is the total number of trees.

3) Synergistic iterative mechanism: BO-GBM synergistic training constitutes the dynamical system:

$$\begin{cases} x^{(k+1)} = \mathcal{B}\left(x^{(k)} | \{\mathcal{J}(x^{(i)})\}_{i=1}^{k}\right) \\ \mathcal{J}(x^{(k)}) = \mathcal{T}_{\mathrm{GBM}}(\mathcal{D}_{\mathrm{train}}, \mathcal{D}_{\mathrm{val}}, x^{(k)}) \end{cases} \quad (31)$$

where, $\mathcal{B}$ is the Bayesian optimization operator and $\mathcal{T}_{\mathrm{GBM}}$ is the GBM training evaluation operator [24].

The sequence $\{\mathcal{J}(x^{(k)})\}$ converges with probability 1 when the hyperparameter space $\mathcal{X}$ is tight and the objective function $\mathcal{J}$Lipschitz is continuous:

$$\lim_{k\to\infty} \mathbb{P}\left(\mathcal{J}(x^{(k)}) - \mathcal{J}^* < \epsilon\right) = 1, \quad \forall \epsilon > 0 \quad (32)$$

4) Computational acceleration strategy

a) Agent model warm-up: Pre-training the auxiliary Gaussian process using historical project data $\mathcal{GP}_0$:

$$\mu_0(x) = \mathbb{E}_{\mathcal{D}_{\mathrm{hist}}}[\mathcal{J}(x)], \quad k_0(x, x') = \mathrm{Cov}_{\mathcal{D}_{\mathrm{hist}}}(\mathcal{J}(x), \mathcal{J}(x')) \quad (33)$$

The initial objective function is estimated as:

$$\mathcal{J}^{(1)}(x) \sim \mathcal{GP}\left(\mu_0(x), \rho k_0(x, x') + k(x, x')\right), \quad \rho = 0.3 (34)$$

b) Gradient-sensitive sampling: Introducing gradient information in EI optimization:

$$x_{\mathrm{next}} = \mathrm{argmax}[\mathrm{EI}(x) + \lambda \parallel \nabla_x \mu(x) \parallel_2] \quad (35)$$

λ Balancing exploration and exploitation to accelerate local convergence [25].

c) Partial evaluation mechanisms: When $\sigma(x) > \sigma_{\mathrm{thres}}$, sub-sampled data are used for evaluation:

$$\tilde{\mathcal{J}}(x) = \mathcal{J}(x; \mathcal{D}_{\mathrm{sub}}), \quad |\mathcal{D}_{\mathrm{sub}}| = \min\left(500, \frac{n}{\sigma(x)/\sigma_{\max}}\right) \quad (36)$$

III. Engineering Economic Risk Management Platform Architecture Design

A. Overall Platform Architecture Design

The engineering economic risk management platform adopts vertical layered architecture, with clear functions and interfaces of each layer to ensure high availability and scalability of the system (see Fig. 3).

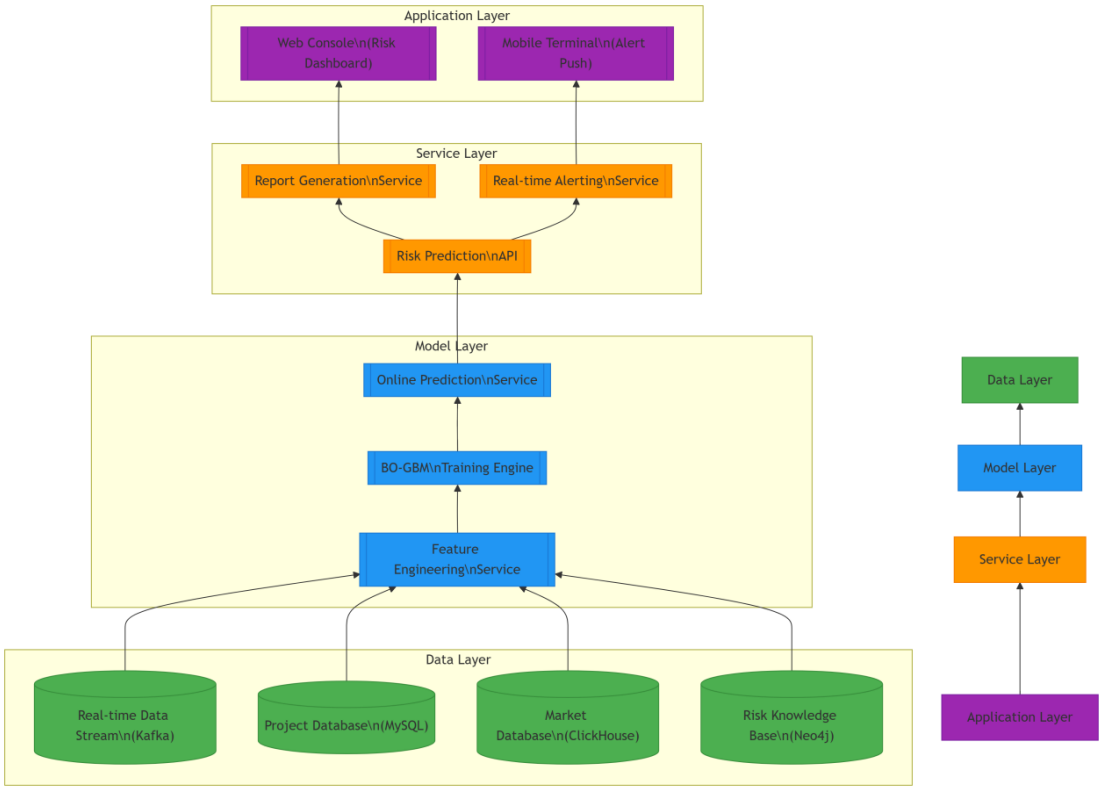


Fig. 3. Platform architecture diagram.

The whole system architecture consists of four distinct layers. The data layer is the foundation and contains both static and dynamic data sources. Storage includes a project database (storing core project data such as contracts, schedules, and costs), a market database (integrating time-series data such as building material prices, exchange rates, and interest rates), and a risk knowledge base (accumulating historical risk cases and response scenarios). Real-time data processing is handled by a Kafka streaming engine that feeds into the business system data and processes it at a rate of over 10,000 data points per minute. The model layer builds on this data foundation and contains the core machine learning components. The BO-GBM training engine automatically optimizes model parameters. The feature engineering service computes 28 real-time risk features. Finally, the Online Prediction Service deploys optimized risk prediction models to ensure that single prediction latency is kept below 300 milliseconds, and the models are updated through daily

automated training. The service layer exposes the system's forecasting capabilities through a set of core services. The Risk Prediction API uses gRPC to calculate project risk probabilities in real time. The real-time alert service uses WebSockets to trigger multi-level alerts. The report generation service creates risk assessment reports automatically via a REST API. To ensure resilience, the service implements automatic meltdown and switches to a backup service when the error rate exceeds 5%. Load balancing supports more than 500 concurrent requests. Finally, the application layer provides the user interface and access control. The web console provides real-time risk heat maps, risk factor analysis tools, and alert threshold configuration options. Mobile terminals provide instant risk alerts, project status queries, and access to emergency response channels. RBAC (role-based access control) is implemented for rights management, and two-factor authentication is used to enhance security.

B. Core Functional Module Design

The engineering economic risk management platform contains five core functional modules, which together achieve the closed-loop management of the whole process of risk prediction, assessment, early warning, and decision-making support. The design of each module follows the principle of 'high cohesion and low coupling', and works together through standardized API interfaces.

1) Data integration and management module: The platform is positioned as a data hub, providing unified access and governance for various heterogeneous data sources. Key design elements include: Multi-source data access: This covers a wide range of systems, including obtaining progress and cost data from project management systems (e.g. Primavera P6) via ODBC, synchronizing payment and settlement information from financial systems [27] (e.g. SAP) via RFC protocols, retrieving building material prices and exchange rate indices on a regular basis from market databases (e.g. Wind) via APIs, and collecting data on the operation of field equipment using IoT sensors. IoT sensors to collect data on the operation of equipment in the field. A key component is the intelligent data pipeline, which contains cleansing rules (e.g., missing value filling and outlier correction using the $\pm 3\sigma$ principle) and transformation logic that automatically generates 28-dimensional feature vectors based on the feature system described earlier. The storage strategy is tiered for optimal performance and cost-effectiveness: hot data is cached in Redis (with response times of less than 50 milliseconds), warm data is stored in ClickHouse columnar format, and cold data is archived in HDFS. Finally, comprehensive metadata management provides data lineage tracking, recording field-level source and transformation history, and quality monitoring dashboards that display data completeness and timeliness metrics in real time (see Fig. 4).

2) BO-GBM model management module: Functional positioning is the core of model full life cycle management. The key design includes: training workflow: multiple triggering mechanisms are available, including timed triggering (2:00 am every day), data drift warning triggering (when the KL dispersion of the feature distribution is > 0.1), and manual triggering. Spark ML is used for parallel processing of 100 GB data for distributed training. Each model version has version metadata, including snapshots of the training data, hyperparameters, and evaluation metrics. Grey scale releases validate the performance of new model versions by introducing 10% traffic. Additionally, the Parameter Configuration Centre provides a BO optimization space that allows for visual adjustment of hyperparameter ranges and supports manual intervention of feature weights to adjust feature importance assignments (see Fig. 5).

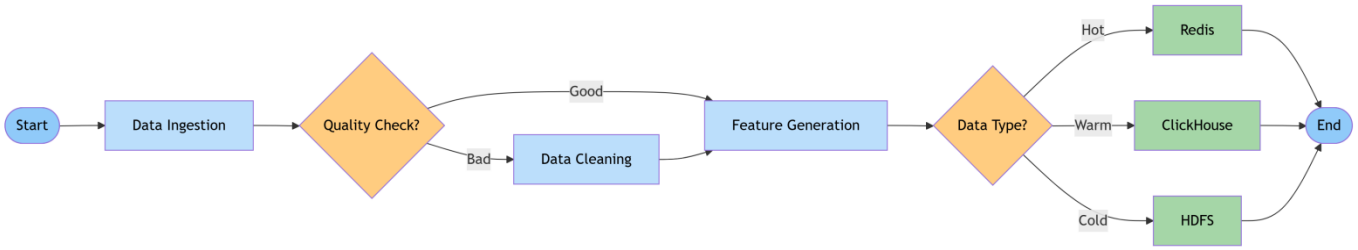


Fig. 4. Intelligent data pipeline flowchart.

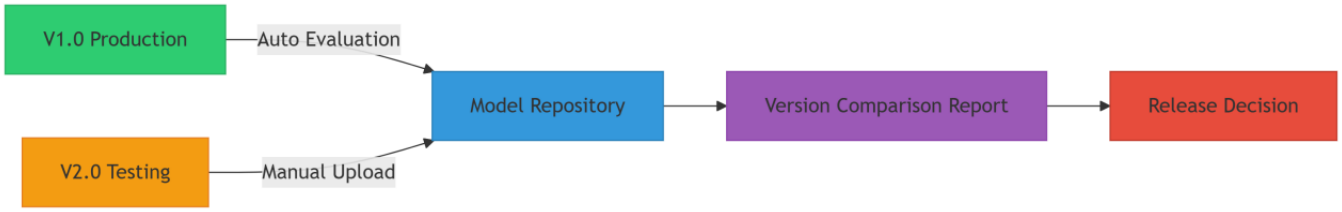


Fig. 5. Version control system.

IV. Platform Core Characteristics Analysis and Verification

A. Validation of the BO-GBM Model's Prediction Performance Characteristics

This experiment verifies the performance advantages of the BO-GBM model in engineering economic risk prediction tasks through rigorous comparison tests. The experimental design focuses on three core questions:

Does BO-GBM significantly outperform mainstream baseline models?

How effective is Bayesian optimization in improving model robustness?

How much does the feature engineering system contribute to the prediction accuracy?

1) Experimental design: The dataset is derived from the data of 327 projects of a large engineering group between 2018 and 2023. The dataset is divided into a training set and a test set, where the training set contains 256 projects from 2018-2022 and the test set contains 71 projects from 2023. Each project is represented by 28 features, and the labels are dichotomized to indicate whether a significant economic risk has occurred (see Table III).

Evaluation metrics include AUC, F1-score, and recall for classification tasks and RMSE and MAPE for regression tasks (cost bias prediction).

2) Performance comparison metrics: This experiment evaluated the classification performance of the five models on 71 independent test items (risk occurrence prediction). All models use the same training set (256 items) and feature set (28 dimensions), and the evaluation metrics include AUC (area under the curve), F1-score (the reconciled average of precision and recall), and recall (the proportion of actual risks correctly identified). The test set consists of 18 high-risk projects (where a significant economic risk actually occurs) and 53 low-risk projects completed in 2023 (see Fig. 6).

The BO-GBM model achieves an excellent performance of 0.927 on the AUC metric, which is significantly higher than the standard GBM (0.865) and Random Forest (0.832). In terms of F1-score, BO-GBM leads the other models with 0.892, which is 7.3% higher than the standard GBM (0.831). The recall metric shows that BO-GBM identifies 91.3% of actual risky items, 17.5 percentage points higher than logistic regression (73.8%). These data demonstrate that Bayesian optimization significantly improves the discriminative power of the GBM model, especially in the identification of high-risk items.

TABLE III Comparative Models

Model type	Parameter Settings	Implementation library
BO-GBM	Tree Depth = 6, Learning Rate = 0.1 (BO Optimization)	LightGBM
Standard GBM	Tree Depth = 6, Learning Rate = 0.1	Scikit-learn
Random Forest (RF)	No. of Trees = 200, Depth = 10	Scikit-learn
SVM	Kernel Function = RBF, C = 1.0	Scikit-learn
Logistic Regression (LR)	Regularization Strength = 1.0	Scikit-learn

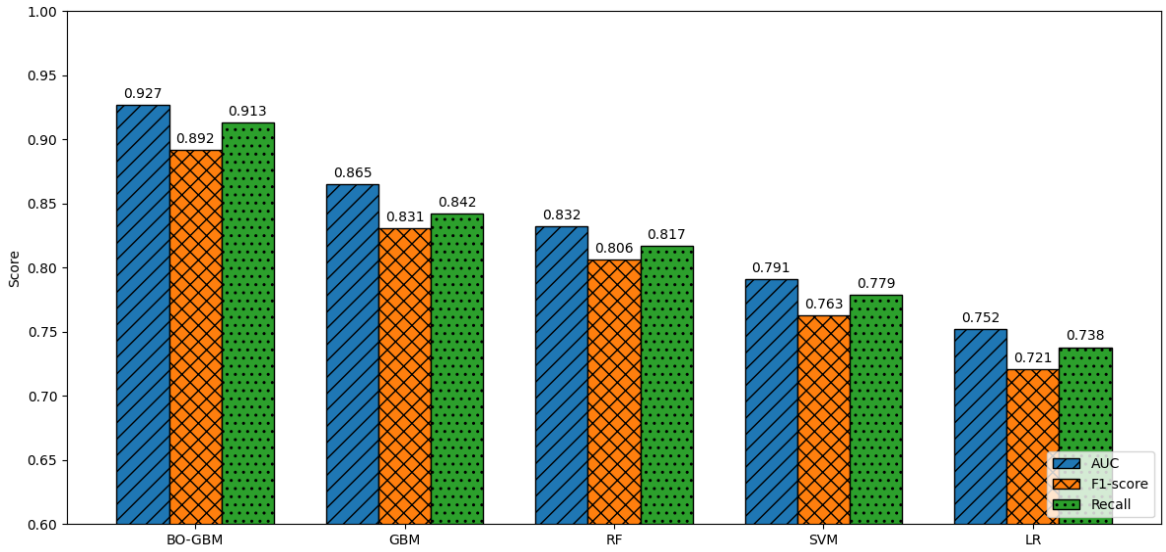


Fig. 6. Comparison of classification model performance.

This experiment evaluated the accuracy of each model in a cost deviation prediction task (regression problem). RMSE (Root Mean Square Error, in millions of dollars) and MAPE (Mean Absolute Percentage Error) were used as assessment metrics. The test set contained actual cost deviation data for 71 projects, with deviations ranging from -35% (savings) to +82% (overruns), and an average absolute deviation of $287,000 (see Fig. 7).

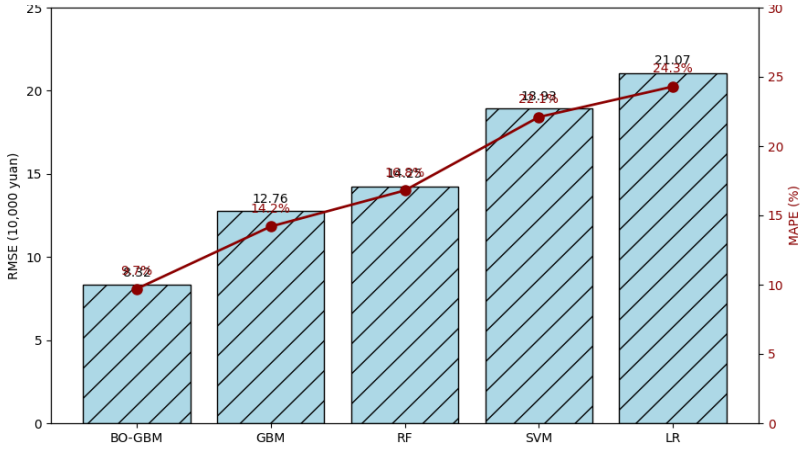


Fig. 7. Comparison of cost bias prediction errors.

The RMSE of BO-GBM is $83,200, which is 34.8% lower than the standard GBM ($127,600) and 60.5% lower than the worst performing logistic regression ($210,700). On the MAPE metric, BO-GBM's prediction error was only 9.7%, breaking the 10% engineering management accuracy threshold for the first time. Notably, when the cost deviation exceeds 30%, the prediction error of BO-GBM (MAPE = 11.2%) is still significantly lower than that of other models (17.8% for standard GBM), indicating that it still maintains high accuracy under extreme risk scenarios.

3) Result analysis: To verify the model's ability to resist interference, Gaussian noise (mean 0, standard deviation from 0% to 20%) was added to the test set of features. The experiment was repeated 50 times at each noise level, and the change in the mean AUC value was recorded. The range of noise addition covers all 28 features to simulate the data acquisition errors in real applications (see Fig. 8).

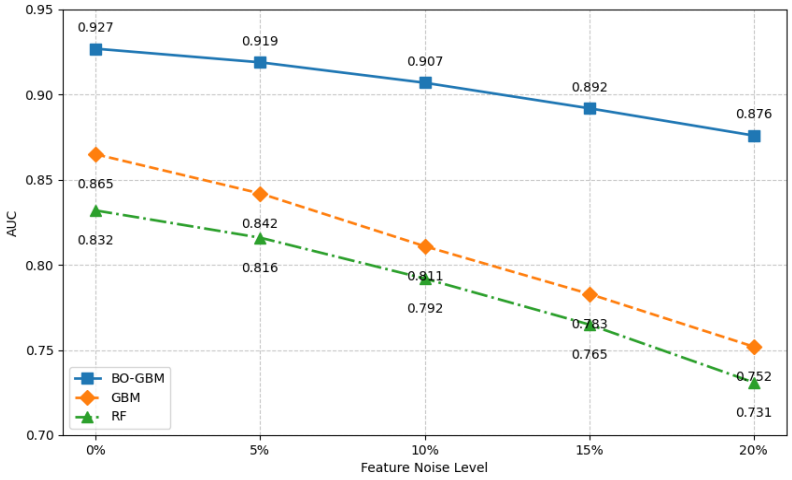


Fig. 8. Model robustness under feature perturbation.

When the feature noise reaches 20%, the AUC of BO-GBM decreases from 0.927 to 0.876 (a decrease of 5.5%), while the standard GBM decreases from 0.865 to 0.752 (a decrease of 13.1%). At a 10% noise level, the AUC of BO-GBM stays above 0.907, which is significantly higher than that of the standard GBM at 0.811. Random Forest performs close to BO-GBM (0.816 vs. 0.919) at low noise (5%), but the AUC decreases to 0.731 at high noise (20%), widening the gap to 14.5 percentage points. This indicates that BO optimization effectively improves the model noise immunity through hyperparameter tuning.

The contribution of each feature to the prediction of the BO-GBM model was quantified using the SHAP value method. The absolute mean of SHAP based on all samples in the test set was calculated, and the five features with the highest contribution were selected. The SHAP value indicates the average magnitude of the effect of feature changes on the model output (risk probability).

The Cost Performance Index (CPI) tops the list with a contribution of 0.218, proving that cost control failure is the strongest risk signal. Building material volatility (0.195) and contract type (0.172) rank second and third, and the sum of their contributions (0.367) exceeds the CPI, reflecting the key role of external markets and contract design. Progress deviation (0.141) and cash flow gap (0.103) constitute the second tier of risk factors. The total contribution of TOP5

features reaches 0.829, covering the three core dimensions of engineering economic risk: cost control, external environment and project execution (see Fig. 9).

B. Platform Functional Characteristics and Efficiency Verification

1) Risk identification efficiency: Fifty projects implemented by a large infrastructure company in 2023 are selected for risk assessment using both traditional manual assessment and automated identification by the platform. Record the time consumed for the whole process from data preparation to the output of the risk assessment report. Manual assessment is performed by a team of three senior risk analysts.

Platform's risk identification efficiency is significantly better than manual methods. For large-sized projects, the identification time is reduced from 42.6 hours to 0.52 hours, an 81.9-fold increase in efficiency. For medium-sized projects, the identification time is reduced from 25.7 hours to 0.38 hours, an increase in efficiency of 67.6 times. Small projects are processed through the platform in 0.25 hours, 49.2 times faster than manual processing (12.3 hours). The logarithmic coordinates show that the platform processing time is basically not affected by the project scale, which verifies the elasticity and scalability of the architecture (see Fig. 10).

2) Predicting response time: The experimental design includes simulating different concurrency scenarios in the load test environment: single project prediction (1000 consecutive requests), multi-project batch prediction (10-100 projects/batch), and high concurrency scenarios (50-500 concurrent users). The platform's response time (P95) and resource consumption (CPU/memory) were recorded. The test environment is an 8-core 16GB cloud server [26].

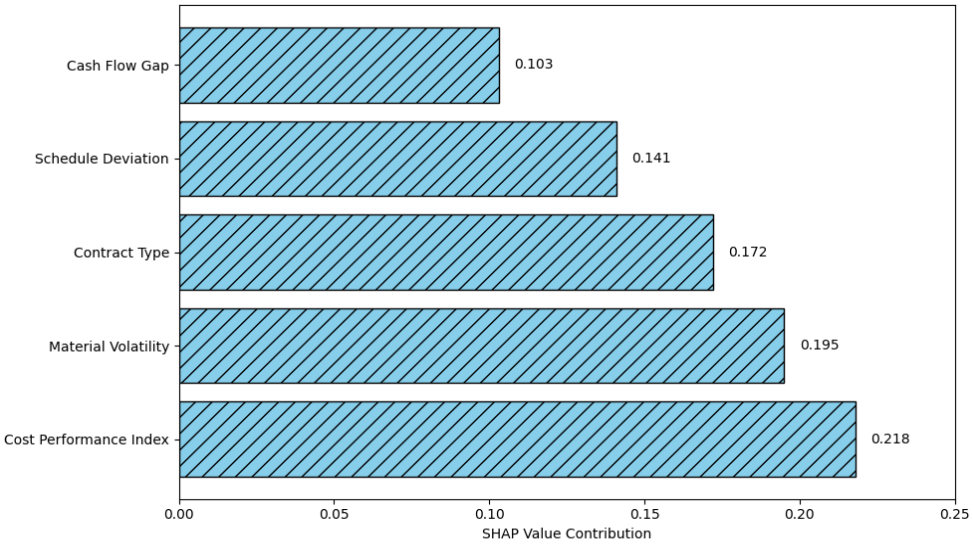


Fig. 9. TOP5 risk characteristics contribution.

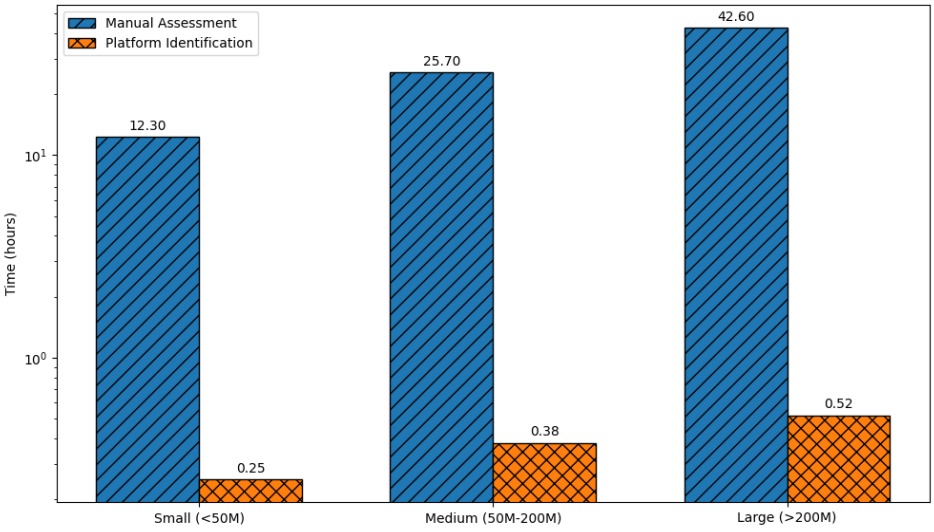


Fig. 10. Comparison of risk identification time for projects of different sizes.

The average response time for a single project prediction is 127 ms, which meets the real-time decision-making requirements. Batch prediction of 100 projects takes 256 ms, with a processing speed of 390 projects per second. In the 500 concurrent users scenario, the P95 response time was 427 milliseconds, still below the engineering threshold of 500 milliseconds. Response times for all test scenarios were below the industry best practice standard of 300 milliseconds. Resource monitoring shows that with 500 concurrent users, CPU utilization is 78% and memory usage is 68%, indicating that the system still has room for further expansion (see Fig. 11).

3) Early warning accuracy verification: The platform was used to backtest the data of 120 projects that had been completed in the past. The specific steps are: input feature data according to the actual project progress nodes, record the warning signals issued by the platform and the time, and compare the warning signals with the time and type of actual risk occurrence. Evaluation indicators include: warning accuracy rate, false alarm rate, and missed alarm rate.

The platform's early warning accuracy rate is 72.5% (87/120), of which 92.3% for high-risk projects (L4-L5). False alarm rate is 7.5% (9/120), which mainly occurs in scenarios of sudden policy changes, such as new environmental regulations. False alarms were 5.0% (6/120), mainly in extreme cases where the schedule was compressed by more than 30%. Overall, the platform made a correct judgment (accurate warning + correct no warning) in 87.5% of cases. The average lead time for high-risk warnings was 28 days for cost risks and 42 days for schedule risks, which meets the emergency response time requirements (see Fig. 12).

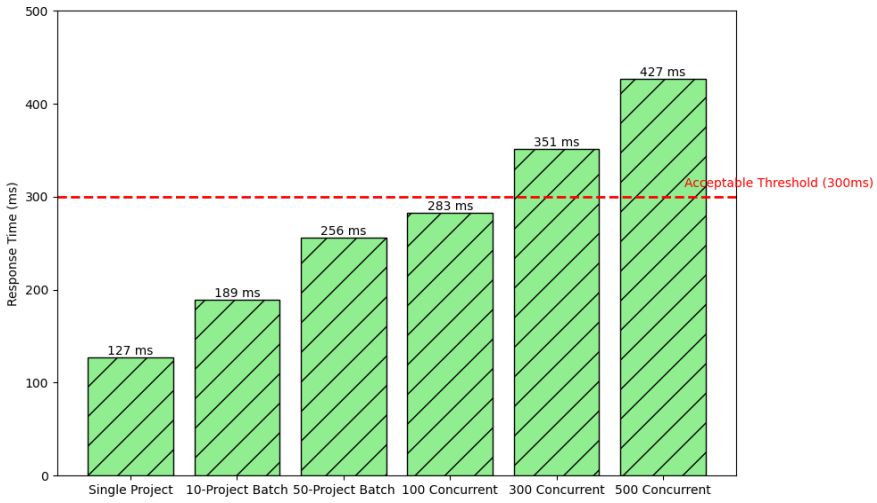


Fig. 11. Risk prediction response time in different scenarios.

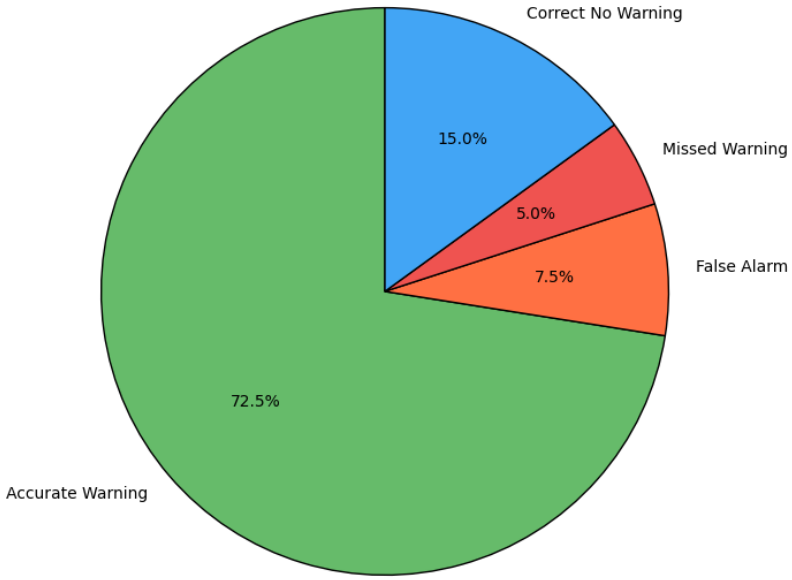


Fig. 12. Analysis of the accuracy of the platform's early warning.

V. Results and Discussion

This section presents a comprehensive evaluation of the proposed BO-GBM model's predictive performance and an analysis of the platform's functional characteristics and efficiency, followed by a discussion of the implications of these findings.

A. Validation of BO-GBM Model's Prediction Performance

The experiment verifies the performance advantages of the BO-GBM model through rigorous comparison tests on a dataset comprising 327 projects from a large engineering group (2018 to 2023), split into training (2018 to 2022, 256 projects) and test (2023, 71 projects) sets. Each project is represented by the 28 features described previously.

1) Experimental setup and baselines: The BO-GBM model was compared against several mainstream baseline models, as detailed in Table IV.

TABLE IV Comparative Models and Parameter Settings

Model Type	Parameter Settings	Implementation Library
BO-GBM	Tree Depth=6, Learning Rate=0.1 (BO Optimized)	LightGBM
Standard GBM	Tree Depth=6, Learning Rate=0.1	Scikit-learn
Random Forest (RF)	No. of Trees=200, Depth=10	Scikit-learn
SVM	Kernel Function=RBF, C=1.0	Scikit-learn
Logistic Regression (LR)	Regularization Strength=1.0	Scikit-learn

2) Performance comparison

a) Classification performance: Fig. 6 shows the performance on the risk occurrence prediction task (18 high-risk, 53 low-risk projects). The BO-GBM model achieved an AUC of 0.927, significantly outperforming the standard GBM (0.865) and Random Forest (0.832). Its F1-score of 0.892 was 7.3% higher than the standard GBM. Crucially, the recall rate of BO-GBM reached 91.3%, meaning it identified over 91% of actual high-risk projects, which is 17.5 percentage points higher than Logistic Regression (73.8%). This demonstrates that Bayesian optimization significantly enhances the model's ability to discriminate, particularly in identifying critical risks.

b) Regression performance (cost deviation): Fig. 7 compares the models on predicting cost deviation. The BO-GBM model achieved an RMSE of 83,200 RMB, which is 34.8% lower than the standard GBM (127,600 RMB) and 60.5% lower than Logistic Regression (210,700 RMB). Its MAPE was only 9.7%, breaking the 10% accuracy threshold often sought in engineering management. Notably, even for extreme cost overruns (>30%), BO-GBM maintained a lower error (MAPE=11.2%) compared to other models (e.g., 17.8% for standard GBM).

3) Robustness analysis: To test the model's resistance to data noise, Gaussian noise (mean 0, standard deviation from 0% to 20%) was added to the test features. As shown in Fig. 8, when feature noise reached 20%, the AUC of BO-GBM decreased by only 5.5% (from 0.927 to 0.876), whereas the standard GBM decreased by 13.1% (0.865 to 0.752). This indicates that the hyperparameters found by BO contribute to a more robust model that is less sensitive to data perturbations, a critical property for real-world applications where data quality can vary.

4) Feature importance analysis: Using SHAP values, we quantified the contribution of each feature. Fig. 9 reveals that the Cost Performance Index (CPI) was the most influential feature (contribution 0.218), confirming failed cost control as the primary risk signal. Building material volatility (0.195) and contract type (0.172) were the next most important, highlighting the significant role of external markets and contractual design. Schedule deviation (0.141) and cash flow gap (0.103) formed a secondary tier. The combined contribution of the top five features was 0.829, effectively capturing the core dimensions of engineering economic risk.

B. Platform Functional Characteristics and Efficiency Verification

1) Risk identification efficiency: We compared the platform's automated risk identification against traditional manual assessment by a team of three senior analysts across 50 projects of different sizes in 2023. The results, depicted in Fig. 10, show a dramatic efficiency improvement. For large projects, identification time was reduced from 42.6 hours to 0.52 hours—an 81.9-fold increase. The platform's processing time remained relatively constant regardless of project scale, demonstrating its scalability and architectural elasticity.

2) Prediction response time: Load testing under various scenarios (Fig. 11) confirmed the platform's real-time capability. The average response time for a single prediction was 127 ms. Batch prediction of 100 projects took 256 ms ($\approx$ 390 projects/second). Under a high load of 500 concurrent users, the P95 response time was 427 ms, remaining below the 500 ms engineering threshold. Resource utilization (CPU 78%, Memory 68%) under this load indicated potential for further scaling.

3) Early warning accuracy: A backtest was conducted on 120 completed projects. The platform achieved an overall warning accuracy of 72.5% (87/120), with accuracy for high-risk projects (L4-L5) exceeding 92.3%. The false alarm rate was 7.5%, primarily triggered by unforeseen policy changes, and the missed alarm rate was 5.0%, mainly occurring in cases of extreme schedule compression (>30%). Overall, the platform made correct judgments (accurate warning + correct non-warning) in 87.5% of cases. The average lead time for high-risk warnings was 28 days for cost risks and 42 days for schedule risks, providing sufficient time for proactive mitigation.

C. Discussion

The results strongly support the effectiveness of the proposed BO-GBM model and the associated platform. The significant performance gains over baseline models, particularly in recall and robustness, underscore the value of using Bayesian Optimization for hyperparameter tuning in this

domain. The high contribution of the engineered features validates the comprehensiveness of the four-dimensional feature system. The platform's operational metrics confirm its practical utility, offering order-of-magnitude efficiency gains in risk identification and reliable, real-time predictions suitable for large-scale, concurrent use.

The primary limitation observed was related to warning errors. False alarms were often linked to "black swan" events like sudden policy shifts, which are not captured by historical feature data. Missed alarms occurred under extreme project conditions, suggesting potential model performance boundaries or the need for even more specialized features for such edge cases. These points inform valuable directions for future work.

VI. Conclusion

In this study, a platform for engineering economic risk management based on the BO-GBM model is successfully constructed, which significantly improves the prediction accuracy and control efficiency of engineering economic risks by integrating Bayesian optimization and gradient boosting machine techniques. The core achievements can be summarized in the following three aspects:

1) Excellent performance of BO-GBM model: By adopting Bayesian optimization to adaptively adjust the GBM hyperparameters, the model breaks through the efficiency bottleneck of traditional tuning methods and significantly enhances the model's robustness. Experimental results show that in the risk prediction task, the model AUC reaches 0.927, with a recall rate of 91.3%, which is 17.5 percentage points higher than that of the standard GBM; in cost deviation prediction, the RMSE is reduced to 83.2 thousand yuan, and the MAPE is only 9.7%, which is more than 30% lower than the mainstream model even in the extreme overrun scenarios; in the face of the featured noise interference of 20%, the model performance degradation is less than 20%. Even under extreme overspending scenarios, the error is still lower than mainstream models by more than 30%; in the face of 20% of characteristic noise interference, the model's performance degradation is less than 6%, which verifies its strong anti-interference ability.

2) Breakthrough in risk characteristics system and platform effectiveness: The four-dimensional 28-feature system (project ontology, market environment, execution process, and risk correlation) constructed systematically quantifies the risk-driving mechanism, with the contribution of key features exceeding 82%. Among them, cost performance index, building material volatility, and contract type constitute the core risk signals. The platform adopts a layered architecture (data layer, model layer, service layer, application layer), supporting millisecond response with single prediction latency of less than 127 milliseconds, and P95 response time of 427 milliseconds with 500 concurrent users. In addition, the efficiency of risk identification has been improved by an order of magnitude, and the evaluation time of large projects has been compressed from 42.6 hours to 0.52 hours, which is 81.9 times more efficient.

3) Precise and reliable early warning mechanism: The platform realizes closed-loop risk management for the whole process. The accuracy rate of early warning reaches 72.5%, and the accuracy rate of high-risk early warning exceeds 92%. The advance warning period for cost risk and schedule risk reaches 28 days and 42 days, respectively, which meets the demand for emergency response. The false alarm rate and omission rate are controlled at 7.5% and 5.0% respectively, which are significantly better than the manual assessment mode.

Building upon current research findings and limitations, future studies will focus on advancing three key areas of exploration. Regarding model optimization, hybrid time-series models incorporating LSTM or Transformer architectures will be developed to more accurately capture the dynamic evolution of project risks. Concurrently, the BO-GBM framework will be expanded to enable multi-task collaborative prediction of cost, schedule, and safety risks. Regarding data dimensions, plans include introducing natural language processing techniques to extract latent risk signals from unstructured data such as textual reports, alongside constructing adaptive feature selection mechanisms to accommodate requirements across different project phases. At the platform functionality level, efforts will strengthen visualization capabilities to intuitively display model decision-making logic, while integrating reinforcement learning technologies to enable dynamic, intelligent recommendations for risk mitigation strategies.

Funding

Chongqing Municipal Education Commission Science and Technology Research Project: Key Technologies for Structural Health Monitoring Data Reconstruction Based on Deep Models (KJQN202301803).

References

[1] Landry M, Erlinger T P, Patschke D, et al. Probabilistic gradient boosting machines for GEFCom2014 wind forecasting. International Journal of Forecasting, 2016, 32(3): 1061-1066.

[2] Yoon J. Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. Computational Economics, 2021, 57(1): 247-265.

[3] Singh U, Rizwan M, Alaraj M, et al. A machine learning-based gradient boosting regression approach for wind power production forecasting: A step towards smart grid environments. Energies, 2021, 14(16): 5196.

[4] Lloyd J R. GEFCom2012 hierarchical load forecasting: Gradient boosting machines and Gaussian processes. International Journal of Forecasting, 2014, 30(2): 369-374.

[5] Taieb S B, Hyndman R J. A gradient boosting approach to the Kaggle load forecasting competition. International journal of forecasting, 2014, 30(2): 382-394.

[6] Xenochristou M, Hutton C, Hofman J, et al. Water demand forecasting accuracy and influencing factors at different spatial scales using a gradient boosting machine. Water Resources Research, 2020, 56(8): e2019WR026304.

[7] Noorunnahar M, Chowdhury A H, Mila F A. A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh. PloS one, 2023, 18(3): e0283452.

[8] Callens A, Morichon D, Abadie S, et al. Using Random forest and Gradient boosting trees to improve wave forecast at a specific location.

Applied Ocean Research, 2020, 104: 102339.

[9] Park S, Jung S, Lee J, et al. A short-term forecasting of wind power outputs based on gradient boosting regression tree algorithms. Energies, 2023, 16(3): 1132.

[10] Panarese A, Settanni G, Vitti V, et al. Develo** and preliminary testing of a machine learning-based platform for sales forecasting using a gradient boosting approach. Applied Sciences, 2022, 12(21): 11054.

[11] Du L, Gao R, Suganthan P N, et al. Bayesian optimization based dynamic ensemble for time series forecasting. Information Sciences, 2022, 591: 155-175.

[12] Sultana N, Hossain S M Z, Almuhaini S H, et al. Bayesian optimization algorithm-based statistical and machine learning approaches for forecasting short-term electricity demand. Energies, 2022, 15(9): 3425.

[13] Cheng H, Ding X, Zhou W, et al. A hybrid electricity price forecasting model with Bayesian optimization for German energy exchange. International Journal of Electrical Power & Energy Systems, 2019, 110: 653-666.

[14] Liwei T, Li F, Yu S, et al. Forecast of LSTM-XGBoost in Stock Price Based on Bayesian Optimization. Intelligent Automation & Soft Computing, 2021, 29(3).

[15] Zulfiqar M, Gamage K A A, Kamran M, et al. Hyperparameter optimization of bayesian neural network using bayesian optimization and intelligent feature engineering for load forecasting. Sensors, 2022, 22(12): 4446.

[16] He F, Zhou J, Feng Z, et al. A hybrid short-term load forecasting model based on variational mode decomposition and long short-term memory networks considering relevant factors with Bayesian optimization algorithm. Applied energy, 2019, 237: 103-116.

[17] Habtemariam E T, Kekeba K, Martínez-Ballesteros M, et al. A Bayesian optimization-based LSTM model for wind power forecasting in the Adama district, Ethiopia. Energies, 2023, 16(5): 2317.

[18] Abbasimehr H, Paki R. Prediction of COVID-19 confirmed cases combining deep learning methods and Bayesian optimization. Chaos, Solitons & Fractals, 2021, 142: 110511.

[19] Aslam M, Lee S J, Khang S H, et al. Two-stage attention over LSTM with Bayesian optimization for day-ahead solar power forecasting. IEEE Access, 2021, 9: 107387-107398.

[20] X B, Zheng W Z, Kong J L, et al. Deep-learning forecasting method for electric power load via attention-based encoder-decoder with bayesian optimization. Energies, 2021, 14(6): 1596.

[21] Nair S K, Botros D, Chakravarti S, et al. Predictors of surgical site infection in glioblastoma patients undergoing craniotomy for tumor resection. Journal of neurosurgery, 2022, 138(5): 1227-1234.

[22] Bidyuk P, Prosyankina-Zharova T, Terentiev O, et al. Adaptive modelling for forecasting economic and financial risks under uncertainty conditions in terms of the economic crisis and social threats. Технологический аудит и резервы производства, 2018, 4(2(42)): 4-10.

[23] Gorokhovatskyi V, Sergienko O, Sosnov I, et al. Risk assessment of innovative projects: Development of forecasting models. CEUR Workshop Proc. 2021, 2927(2020): 18-37.

[24] Danilov V Y, Gozhyj O P, Kalinina I O, et al. Adaptive forecasting and financial risk estimation. System research and information technologies, 2020 (1): 34-53.

[25] Zhao Y, Liu X, Zhao Y. Forecast for construction engineering risk based on fuzzy sets and systems theory. Systems Engineering Procedia, 2011, 1: 156-161.

[26] Bhardwaj A, Sharma A, Raj D, et al. Dynamic and Scalable Privacy-Preserving Group Data Sharing in Secure Cloud Computing. Advanced Cyber Security Techniques for Data, Blockchain, IoT, and Network Protection. IGI Global Scientific Publishing, 2025: 89-122.

[27] Oko-Odion C, Angela O. Risk management frameworks for financial institutions in a rapidly changing economic landscape. International Journal of Science Research Archive, 2025, 14(1): 1182-1204.

[28] Wuni I Y. Develo, A multidimensional risk assessment model for sustainable construction projects. Engineering, Construction and Architectural Management, 2025, 32(6): 4155-4173.

Adaptive Virtual Machine Consolidation Based on Autoformer and Enhanced Double Q-Network for Energy-Efficient Cloud Data Center

Kaiqi Zhang, Youbo Lyu, Dequan Zheng*, Yanping Chen, Jianshan Xu
School of Computer and Information Engineering, Harbin University of Commerce, Harbin 150028, China

***Abstract*—As the scale of cloud data centers continues to expand, energy consumption has become a critical issue. Virtual machine (VM) consolidation is a key technology for improving resource utilization and reducing energy consumption, yet it remains challenging to effectively balance energy efficiency with service level agreement violations (SLAV) in dynamic cloud environments. This paper proposes an adaptive VM consolidation strategy based on Autoformer and an enhanced dual Q-Network, referred to as AEDQN-VMC. The approach consists of three integrated components: 1) Autoformer-based load detection, which leverages an autocorrelation mechanism to decompose time-series data into multi-scale trend and periodic components; 2) a VM selection method that integrates the Pearson correlation coefficient and migration time to optimize the selection of VMs for migration; and 3) an enhanced dual Q-Network for VM placement, incorporating the upper confidence bound (UCB) and adaptive learning rate (ALR) to improve the exploration-exploitation trade-off. Extensive experiments on real-world cloud workload traces (PlanetLab, Google Cluster, and Alibaba datasets) demonstrate that the proposed method significantly outperforms state-of-the-art benchmarks such as PABFD, AD-VMC, and AMO-VMC. Specifically, it achieves maximum reductions of 46.5% in energy consumption and 74.2% in SLAV rate. Ablation studies further validate the contribution of each component and confirm the synergistic effect of the overall architecture. The results highlight the potential of AEDQN-VMC as an efficient and reliable solution for sustainable cloud data center operations.**

Keywords*—Cloud computing; virtual machine consolidation; load prediction; energy efficiency; deep reinforcement learning; Autoformer

I. Introduction

With the rapid development of cloud computing, data centers have become the core infrastructure supporting various online services, including big data analysis, artificial intelligence, and Internet of Things (IoT) applications. The exponential growth of user demands has driven a significant expansion in the scale of data centers, which typically consist of thousands of physical servers. However, this expansion gives rise to a critical issue: excessive energy consumption, with a considerable portion of energy wasted due to low resource utilization efficiency [1]. High energy consumption not only increases the operational costs for data center operators but also exacerbates carbon emissions, conflicting with global sustainable development and carbon neutrality goals. Therefore, it has become an urgent challenge to reduce energy consumption while ensuring quality of service (QoS) in the field of data center management.

Virtual Machine (VM) consolidation is a core strategy to address the energy consumption issue of data centers. By dynamically migrating VMs between physical machines (PMs), VM consolidation optimizes resource allocation. Through consolidating VMs onto a smaller number of active PMs, idle or underloaded PMs can be shut down or switched to low-power modes, thereby reducing overall energy consumption and improving resource utilization [2, 3]. However, VM consolidation is a complex decision-making process that requires balancing multiple objectives, such as minimizing energy consumption, reducing Service Level Agreement Violation (SLAV), and lowering migration overhead.

The VM consolidation process is divided into three phases: PM state detection, VM selection, and VM placement. Therefore, accurate PM state detection, reasonable VM selection, and optimized VM placement are the three major challenges that must be addressed to solve the VM consolidation problem. To tackle these challenges, Goyal et al. [4] proposed an adaptive multi-objective VM consolidation strategy (AMO-VMC), which aims to optimize resource management in energy-efficient cloud data centers. This strategy fuses the future resource utilization and the historical utilization for identifying overutilized PMs and selecting VMs to be migrated, and a multi-objective heuristic-based adaptive VM placement algorithm is adopted to select the optimal target host. Zeng et al. [5] proposed an adaptive Deep Reinforcement Learning(DRL)-based VM consolidation framework (AD-VMC), whose core components are an influence coefficient(IC)-based VM Selection algorithm (ICVMS) and a prediction aware DRL-based VM placement method(PADRL). By calculating the similarity between VMs and PMs, VMs that have the greatest impact on PM overload are migrated first to quickly relieve load pressure. Additionally, Long Short-Term Memory (LSTM) Neural Network is used to predict future system states to accelerate the convergence of the DRL model, thereby achieving energy-efficient VM placement.

However, existing researches still have significant limitations and struggle to simultaneously meet the requirements of energy efficiency optimization and QoS assurance in dynamic cloud environments. In the PM state detection phase, although workload prediction-based PM state detection methods are widely adopted, and deep learning-based prediction models can capture the temporal dependencies of

*Corresponding author.

workloads [6, 7], their ability to model the complex multi-scale periodicity (e.g., intra-day and intra-week fluctuations) in cloud workloads remains limited. This may lead to misjudgment of overloaded/underloaded hosts. In the VM selection phase, although some studies consider both resource correlation and migration cost [4, 8], they fail to effectively model the non-linear relationship between correlation and cost. In the VM placement phase, researchers have proposed several VM placement schemes based on DRL [9, 10]. However, these schemes adopt fixed exploration probabilities and learning rates, making it difficult to balance exploring new actions and exploiting known optimal actions, thus resulting in the problem of exploration-exploitation imbalance.

To address the aforementioned limitations, an adaptive VM consolidation scheme based on Autoformer and enhanced double Q-Network (AEDQN-VMC) is proposed, which achieves efficient resource management in dynamic environments through synergistic optimization across three phases:

- Autoformer-based PM state detection: Autoformer is employed for workload prediction, which decomposes workload time series into trend and periodic components via an auto-correlation mechanism.
- VM selection based on Pearson correlation coefficient and migration Time: A VM migration impact factor is constructed by fusing the Pearson correlation coefficient and migration time to balance the effect of load reduction and migration cost.
- VM placement based on enhanced double Q-Network: The upper confidence bound (UCB) and adaptive learning rate (ALR) are introduced into the double Q-Network to optimize the exploration-exploitation balance.

The remainder of this paper is organized as follows. Section II presents a classified review of related research on VM consolidation. Section III describes the system model and overall framework. Section IV details the Autoformer-based PM state detection method. Section V proposes the VM selection strategy based on Pearson correlation coefficient and migration time. Section VI introduces the VM placement algorithm based on the enhanced double Q-Network. Section VII evaluates the effectiveness of the proposed strategy through experiments. Section VIII concludes the paper and outlines future research directions.

II. Related Works

As a core strategy for resource management in energy-efficient cloud data centers, VM consolidation aims to significantly reduce system energy consumption by centrally deploying VMs on a minimized cluster of PMs, thereby maximizing physical resource utilization and minimizing the number of active servers. This strategy runs through the entire lifecycle management of VMs. In the initial VM creation phase, VM consolidation manifests as placement decisions based on PM state detection. During the dynamic migration process in the operation phase, VM consolidation involves a three-stage optimization of PM state detection, VM selection, and VM placement. To systematically analyze the inherent logic of the consolidation mechanism, the following sections will sort out existing studies from three core dimensions: PM state detection, VM selection, and VM placement.

A. PM State Detection

Before selecting VMs and making placement decisions, it is necessary to accurately detect the operational status of PMs. Existing PM detection methods can be categorized into two types based on whether they rely on workload trend prediction:

1) Non-prediction-based: This type of method directly determines the PM status using real-time or historical workload data, without the need to predict future workload changes. In [11], PM state detection primarily relies on real-time resource utilization, with CPU and memory as core dimensions. This involves calculating the available resources of PMs and classifying PM states into overloaded, normal, and underloaded by setting thresholds. The study [12] classifies workloads using the split-and-recombine (SAR) algorithm based on real-time monitored resource utilization and workload characteristics, and further considers thermal cycling effects through a thermal model to determine the host state. The study in [13] calculates power consumption across different utilization intervals using a linear interpolation power model, and then evaluates whether a PM can accommodate additional VMs by combining current and historical CPU utilization. The study [14] relies on real-time temperature monitoring to periodically check server temperatures, which dynamically sets a maximum temperature threshold and determines if the server is overloaded when the temperature exceeds this threshold.

2) Load prediction-based: To overcome the response lag of non-prediction-based methods, researchers have introduced workload prediction technique, which enables proactive state identification by forecasting the future resource utilization of PMs. The core value of such methods lies in providing a buffer time window for migration decisions, thereby avoiding SLAVs and reducing the frequency of emergency migrations.

Statistical methods are traditional workload prediction approaches and remain the mainstream modeling techniques for workload prediction to date. The study [15] uses the ARIMA model to analyze the trend of PM resource utilization, predicting future resource utilization to determine whether a PM is overloaded. Reference [16] predicts the future workload of PMs via the exponential smoothing method, which calculates a workload anomaly function by combining the predicted future workload with the current workload, and then determines whether a PM is in an overloaded state. The study [17] infers the future overload probability by analyzing historical host resource usage with the Local Regression algorithm, thereby identifying whether a PM is in an abnormal state.

However, statistical methods have obvious limitations when dealing with complex dynamic workloads. First, most of these methods rely on the assumptions of linearity and stationarity of data. In reality, however, data center workloads

often exhibit nonlinear and non-stationary characteristics, accompanied by sudden fluctuations, which leads to a decline in prediction accuracy. Second, traditional statistical models have poor adaptability to high-dimensional data and struggle to integrate multi-source information, resulting in limited generalization ability in complex scenarios. Third, for workload sequences with long-term dependency characteristics, simple smoothing or regression methods are unable to capture the deep correlation between historical data and future trends, making them prone to lag errors.

To address the limitations of statistical methods, researchers have increasingly adopted deep learning techniques to construct workload prediction models in recent years. The study [10] combines Temporal Convolutional Networks (TCN) with Median Absolute Deviation (MAD) to predict the overload probability of PMs, which reduces unnecessary VM migrations and SLAVs. The study [5] designs an LSTM-based network to forecast future system states, which provides more reasonable environmental states for the DRL model, assisting in more accurate judgments of whether PMs are overloaded or underloaded. The study [18] proposes a virtualized adaptive scheduling algorithm, which uses LSTM to predict future CPU utilization, memory usage, and network bandwidth requirements of VMs, and employs Deep Q-Network (DQN) to achieve adaptive resource scheduling.

Nevertheless, deep learning-based workload prediction confronts several issues, such as high model complexity, substantial training costs, and sensitivity to small-sample data. To further optimize prediction performance, researchers have combined statistical methods with deep learning to construct hybrid workload prediction models. For instance, the study [4] combines the Dynamic Weighted Moving Average (DWMA) and a neural network model to calculate the expected utilization rate of physical machines. The study [19] combines Bollinger Bands and Neural Prophet techniques to predict the future workload trend of hosts.

B. VM Selection

After identifying the source PMs that require adjustment, it is necessary to further select the VMs to be migrated. Based on the criteria used to select VMs from overloaded PMs, relevant studies on VM selection can be categorized into three types: based on PM-VM correlation, based on VM migration overhead, and based on both correlation and migration overhead.

1) PM-VM correlation: The VM selection criteria in this category primarily focus on the degree of association between VMs and PMs in aspects such as resource usage patterns, workload trends, and energy consumption characteristics. By quantifying this association, VMs that have a significant impact on the PM's state are selected. In study [5], VMs are selected based on the Influence Coefficient, which quantifies a VM's contribution to the overload of a PM by calculating the product of the cosine similarity between the VM's and PM's resource usage. VMs with higher influence coefficient exhibit a stronger correlation with the PM's overload and are thus prioritized for migration. In study [20], a load consolidation strategy based on Q-learning selects VMs by learning the load matching patterns between VMs and PMs. Specifically, it prioritizes the migration of VMs that can balance the PM's load, which reflects the correlation between VM and PM load trends. In study [15], VMs are selected based on contribution of the VM to the overload of the PM. VMs with the largest contribution are prioritized for migration, with the goal of eliminating the overload trend of the PM. In [21], with the goal of minimizing energy consumption, VMs that are compatible with the energy consumption characteristics of PMs are selected.

2) VM migration overhead: The VM selection criteria in this category primarily focus on the costs incurred during the migration process (e.g., time, performance loss, energy consumption, and impacts on QoS). By optimizing the selection of VMs, the negative impacts of migration on the system are reduced. In study [11], a fuzzy meta-heuristic algorithm is employed to select VMs whose migration imposes the least impact on QoS. In study [22], the direct criterion for VM selection is the migration time of VMs. The Minimum Migration Time (MMT) algorithm is adopted to prioritize the selection of VMs that can be migrated from overloaded hosts in the shortest time.

3) Both correlation and migration overhead: The criteria for VM selection in this type of research take into account both the correlation between VMs and PMs and the overhead during the migration process. The core is to balance these two aspects to achieve the optimization goal. For example, in study [8], the objective function aims to simultaneously minimize the number of active PMs and the frequency of VM migrations, thereby balancing resource utilization and migration costs. In study [23], a Deep-Q Network is employed to learn the load matching patterns between VMs and PMs, taking into account the impacts of migration on VM performance. It selects VMs that can balance load optimization and low performance loss. In [24], the correlation between VMs and the overload of PMs is determined to select VMs that can alleviate PM overload, taking into account migration time and performance degradation.

C. VM Placement

VM Placement (VMP) refers to the process of allocating newly created VMs or VMs to be migrated to optimal PMs, under the premise of satisfying multi-dimensional resource constraints such as CPU, memory, storage, and network. The rationality of this process directly affects resource utilization, energy consumption, and service quality. In recent years, researchers have proposed a variety of optimization methods to address this problem, which can be categorized into the following three types based on their technical approaches.

1) Heuristic algorithms: These algorithms generate placement strategies based on intuitive experience or predefined rules, with the goal of quickly obtaining feasible solutions, which is suitable for large scale dynamic scenarios. In study [8], a modified First-Fit strategy combined with a backfilling mechanism is adopted. Based on the runtime and priority of VMs, this approach places each VM onto the first

PM that meets the resource constraints. In study [25], the proposed cut-and-solve algorithm is a tree-search heuristic algorithm, which achieves the optimized placement of VMs by iteratively decomposing the problem and solving the sub-problems. In study [24], the VM allocation problem is modeled as a multi-dimensional bin packing problem, the objective of which is to select the minimum number of energy-efficient PMs, thereby to reduce energy consumption and resource waste. In study [26], a two-stage heuristic algorithm is proposed to address the rack-level hot-spot issue. In the first stage, VMs are selected for migration from overloaded hosts within racks identified as hard hot-spots, aiming to reduce the power consumption of these racks. In the second stage, VMs are selected for migration from underloaded hosts; this step serves to reduce the number of active servers, thereby achieving resource consolidation.

Heuristic algorithms offer the advantage of strong real-time performance. However, they struggle to balance multi-objective optimization (e.g., simultaneously minimizing energy consumption and achieving load balancing) and tend to get trapped in local optima.

2) Meta-heuristic algorithms: To address the limitations of heuristic algorithms, meta-heuristic algorithms perform global optimization by simulating natural phenomena (e.g., biological evolution, swarm behavior), which are suitable for complex multi-objective scenarios. The study [11] proposes a hybrid algorithm named IVPTS, which is designed for the collaborative optimization of task scheduling and VM placement in cloud computing environments. This algorithm integrates the Improved Particle Swarm Optimization algorithm with the Fuzzy Resource Management framework to achieve the goals of load balancing and energy conservation. To simultaneously optimize QoS, server energy consumption, and cooling energy consumption, The study [12] proposes a VM Placement strategy named BTVMP, which is based on the Enhanced Simulated Annealing algorithm and integrates a thermal cycling effect model. In study [14], the VM placement involves both heuristic and meta-heuristic methods. In the initial placement phase, a hybrid algorithm combining genetic algorithm and simulated annealing algorithm is adopted to achieve global optimization of the initial VM allocation. In the dynamic migration phase, a dynamic migration strategy based on the greedy algorithm is employed, which selects the server with the highest energy efficiency as the migration target.

3) Reinforcement learning approaches: With the development of machine learning, Reinforcement Learning (RL) and Deep Reinforcement Learning (DRL) have been applied to address the dynamic optimization problem of VM placement. The core of this method lies in modeling the placement process as a Markov Decision Process (MDP), where an agent learns the optimal strategy through interaction with the environment. In study [5], VM placement is implemented based on DRL. An LSTM-based state prediction network provides future environmental states, which accelerates the convergence of the DRL model and enables it to adaptively select the optimal target host for migrated VMs. To address the issue that existing methods optimize sub-steps (such as selecting target PMs for migration and choosing services to be migrated) independently while ignoring the correlations between decisions, Reference [9] proposes a micro-service migration framework based on Multi-Agent RL, which jointly optimizes the three sub-steps of the migration process: selecting source PMs, selecting micro-services, and selecting target PMs. In study [27], a VM migration management strategy based on RL is proposed, which adopts a decentralized operation mode through parallel learning and state/action space aggregation. In study [28], an energy-efficient dynamic consolidation method named MAGNETIC based on the multi-agent Q-Learning algorithm is proposed, in which the agent dynamically selects the optimal power mode to balance energy consumption and QoS.

RL approaches offer the advantage of strong adaptability, enabling them to dynamically adjust strategies to accommodate the uncertainties of cloud environments. However, they face challenges such as high training costs and insufficient convergence stability.

III. System Model and Framework

A. System Model

Suppose a data center consists of *N* PMs denoted as $H = \{H_1, H_2, \dots, H_N\}$, where each PM H_i runs n_i VMs denoted as $V = \{V_1, V_2, \dots, V_{n_i}\}$. VM consolidation refers to the process of reasonably allocating a large number of VMs to PMs, with the core objective of optimizing resource utilization and reducing energy consumption while meeting QoS requirements and resource constraints.

1) Energy Consumption: The energy consumption of a PM is represented using a linear model as follows:

$$E_i = E_{idle} + (E_{max} - E_{idle}) \cdot u_i \quad (1)$$

where E_i denotes the total energy consumption of the PM H_i, E_{idle} represents the base energy consumption in the idle state, E_{max} is the maximum energy consumption under full load, and $u_i \in [0, 1]$ stands for the CPU utilization rate.

The total energy consumption of the entire data center is the sum of the energy consumption of all active PMs:

$$EC = \sum_{i=1}^{N} E_i \cdot \delta_i \quad (2)$$

where $\delta_i \in \{0, 1\}$ indicates the activation state of the PM H_i.

2) Service Level Agreement Violation: Service Level Agreement Violation (SLAV), a key metric for measuring whether VM consolidation affects QoS, refers to the number or proportion of times the QoS specified in the SLA fails to be met. In general, SLAV depends on two metrics: one is the SLAV time of each active host (SLATAH), and the other is the performance degradation caused by migration (PDM). SLAV can be expressed as:

$$SLAV = SLATAH \times PDM \quad (3)$$

where SLATAH represents the proportion of the total operating time that is occupied by SLA violation time in the service agreements of each active host, which is expressed as:

$$SLATAH = \frac{1}{N}\sum_{i=1}^{N}\frac{T_{H_i}^{OVER}}{T_{H_i}^{TOTAL}} \quad (4)$$

where N is the total number of all PMs. For any PM H_i, the total duration during which it violates the SLA due to overload is recorded as $T_{H_i}^{OVER}$, and the total duration during which this host is in an active state is $T_{H_i}^{TOTAL}$.

In study (3), PDM is used to represent the proportion of performance degradation caused by VM migration to its total performance requirements, which is expressed as:

$$PDM = \frac{1}{n}\sum_{j=1}^{n}\frac{S_{d,j}^{OVER}}{R_{d,j}^{TOTAL}} \quad (5)$$

where n represents the number of VMs, d denotes the resource type. For any VM V_j, the amount of performance degradation caused by migration is marked as $S_{d,j}^{OVER}$, and the total resource demand of this VM during its lifecycle is recorded as $R_{d,j}^{TOTAL}$.

B. System Framework

The framework of the VM consolidation system is illustrated in Fig. 1. Adopting a modular design, this framework decomposes the entire VM consolidation process into three key phases, forming a complete closed-loop control system.

- Workload Detection Module: Monitoring the resource utilization of all PMs in real time and outputs three types of PM states.
  - Overloaded PMs: PMs predicted to exceed the safety threshold.
  - Underloaded PMs: Idle PMs whose resource utilization remains below the minimum threshold.
  - Normal PMs: PMs operating within the ideal load range.
- VM Selection Module: For overloaded PMs, VMs are selected to be migrated out. For underloaded PMs, all VMs running on them are marked as to-be-migrated.
- VM Placement Module: The optimal target PM is selected from normal PMs to place the VMs selected in the VM selection module. After executing the migration, idle PMs are shut down to save energy.

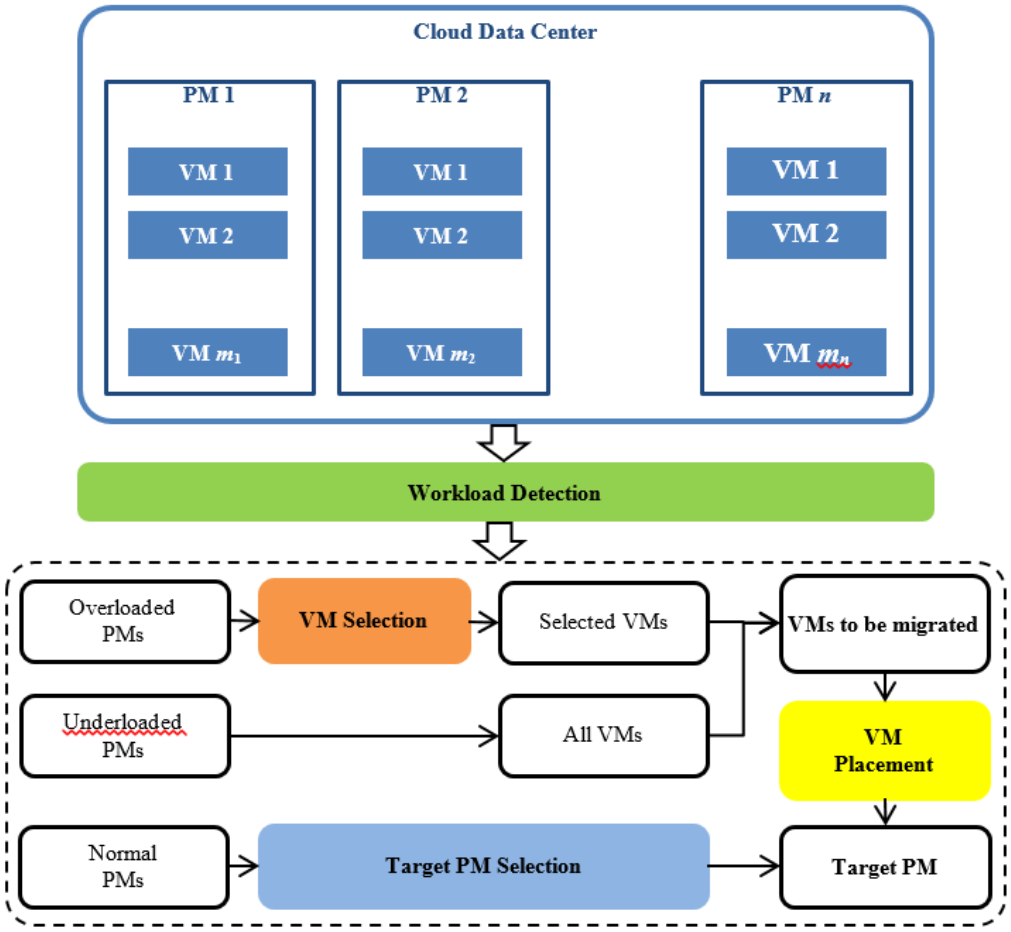


Fig. 1. Framework of VM consolidation.

IV. Workload Detection Based on Autoformer

Cloud workload time-series data typically contain both linear and non-linear components, exhibiting complex time-varying characteristics and significant temporal dependence. In response to this feature, a prediction architecture based on Autoformer is adopted [29], whose core components consist of an encoder and a decoder, as shown in Fig. 2.

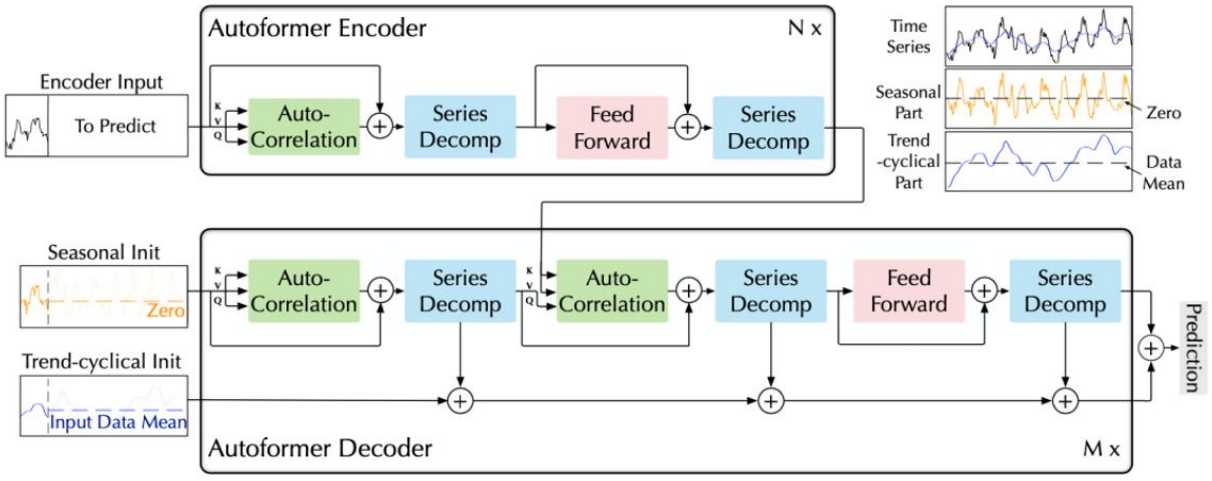


Fig. 2. Cloud workload prediction architecture based on Autoformer [29].

The encoder module takes historical workload time-series as input and performs multi-scale decomposition on the original time-series data through the Auto-Correlation Mechanism. This decomposition decouples the data into two orthogonal components:

- workload trend component which characterizes the long-term evolution law of resources;
- workload periodic component which depicts the inherent periodic fluctuation pattern of the system.

The decoder module adopts a dual-branch structure to process these two components separately. For the periodic component, progressive time-series feature extraction is achieved through cascaded auto-correlation blocks. For the trend component, a cumulative summation operation is used for multi-order trend information aggregation. Each layer of the decoder receives complete time-series segments rather than discrete sampling points. This global perception mechanism significantly enhances the model's ability to model complex workload patterns.

The workload detection process based on Autoformer is shown in Algorithm 1. In the initialization phase, the sets of PMs in overloaded, underloaded, and normal states are initialized as empty sets respectively, and the trained Autoformer model is loaded simultaneously. Next, the algorithm processes each PM in sequence. Line 3 obtains the current utilization rate of the PM by calling the getCurrentUtilization function, while Line 4 uses the getHistoryUtilization function to acquire the historical utilization sequence of the PM according to the window step size. This sequence contains resource usage data within a certain time span and serves as a key input for model prediction. Line 5 inputs the historical utilization sequence into the Autoformer model. Through its encoder, the historical workload time series undergoes multi-scale decomposition to separate the workload trend component and workload periodic component. These two components are then processed separately by the dual-branch structure of the decoder, ultimately outputting the predicted utilization rate. Lines 6–12 achieve accurate classification of PM states by comparing the current utilization rate and predicted utilization rate with the overload threshold and underload threshold. If the current utilization rate or the predicted utilization rate exceeds the overload threshold, the PM is determined to be in an overloaded state. If the predicted utilization rate is lower than the underload threshold and all data in the historical utilization sequence are below the underload threshold, it indicates that the host's resource utilization has been continuously low in the past and will remain low for a period in the future, so the PM is judged to be in an underloaded state. In all other cases, the host is classified as being in a normal state.

Algorithm 1: DetectWorkload

Input: physical machines $H = \{H_1, H_2, \dots, H_N\}$
window steps I
overload and underload threshold th_{over}, th_{under}

Output: overloaded PMs, underloaded PMs, normal PMs

Initialize:

1 $overloaded \leftarrow \emptyset$, $underloaded \leftarrow \emptyset$, $normal \leftarrow \emptyset$
Autoformer ← **load pretrained Autoformer model**

2 **For** each host H_i in H **Do**

3 $current_util \leftarrow$ getCurrentUtilization(H_i)

4 $util_sequence \leftarrow$ getHistoryUtilization(H_i, I)

5 $predicted_util \leftarrow$ ***Autoformer*****.predict(** ***util_sequence*** **)**

6 **If** $current_util > th_{over}$ OR $predicted_util > th_{over}$ **Then:**

7 $overloaded \leftarrow overloaded \cup \{H_i\}$

Else If *predicted_util* < th_{under} AND all(*util* < th_{under} for *util* in *util_sequence*) **Then:**

    underloaded ← *underloaded* ∪ { H_i}

Else:

    normal ← *normal* ∪ { H_i}

End If

End For

return *overloaded*, *underloaded*, *normal*

V. VM Selection Based on Pearson Correlation Coefficient and Migration Time

In cloud computing environments dynamically changing, VM selection strategy is a key link that affects consolidation efficiency. An unreasonable VM selection strategy may lead to low migration efficiency or the risk of secondary overload. Therefore, this section proposes a comprehensive selection strategy that combines the Pearson correlation coefficient and migration time.

The Pearson correlation coefficient is a core metric for measuring the linear correlation between two variables. In the context of VM selection, it can accurately characterize the degree of correlation between the resource utilization of a VM and that of a PM. The equation is as follows:

$$\rho(V,H) = \frac{\sum_{t=1}^{T}(U_V^t-\mu_V)(U_H^t-\mu_H)}{\sqrt{\sum_{t=1}^{T}(U_V^t-\mu_V)^2}\sqrt{\sum_{t=1}^{T}(U_H^t-\mu_H)^2}} \quad (6)$$

where V denotes a VM, H represents a PM, and T stands for the length of the historical time window. U_V^t is the CPU utilization of the VM at time t, and U_H^t is the CPU utilization of the PM at time t. $\mu_V = \frac{1}{T}\sum_{t=1}^{T} U_V^t$ is the mean value of the CPU utilization time series of the VM, while $\mu_H = \frac{1}{T}\sum_{t=1}^{T} U_H^t$ is the mean value of the CPU utilization time series of the PM.

The time overhead for migrating virtual machine V from host H is modeled as:

$$MigTime(V,H) = Mem(V)/BW_{avail}(H) \quad (7)$$

where $\mathrm{Mem(V)}$ is the memory usage of the virtual machine, and $\mathrm{BW_{avail}(H)}$ represents the available network bandwidth of the host.

The VM migration impact factor based on the Pearson correlation coefficient and migration time is defined as follows:

$$IF(V,H) = \rho^2(V,H)/MigTime(V,H) \quad (8)$$

where the square operation strengthens the impact of VMs with high correlation, and the denominator term penalizes the migration of VMs with large memory.

Based on the definition of the VM migration impact factor, a VM selection algorithm that combines the Pearson correlation coefficient and migration time is proposed, as shown in Algorithm 2. This algorithm adopts a greedy strategy, selecting the VM with the largest current IF value for migration until the resource utilization of the overloaded host drops below the safety threshold.

Algorithm 2: SelectVMS

Input: overloaded hosts set $H_{overload}$

Output: migration list M

Initialize: $M \leftarrow \emptyset$

For each host H_i in $H_{overload}$ **Do**

    $V^* \leftarrow \underset{V \in H_i}{\mathrm{argmax}}\, IF(V,H_i)$

    $M \leftarrow M \cup \{V^*\}$

    $H_i \leftarrow H_i \setminus \{V^*\}$

End For

return M

VI. VM Placement Based on Enhanced Dual Q Network

VM placement refers to the reasonable allocation of VMs to PMs, aiming to optimize resource utilization and reduce energy consumption.

A. Problem Model

The VM placement framework based on the improved double Q neural network is illustrated in Fig. 3. The double Q neural network consists of two Q networks with identical structures but different parameter update methods: the online Q network QNet(θ) and the target Q network TargetQNet(θ^-). The online Q network is used to select action $\mathrm{a_t}$ based on the current environmental state $\mathrm{s_t}$, while the target Q network is employed to calculate the target Q-value. This design avoids the overestimation problem that may occur in traditional Q-learning, making the estimation of Q-values more accurate.

1) State representation: The environmental state s_t includes the CPU utilization information of each PM and the CPU demand information of the VM, which is defined as $s_t = [u_1^t, u_2^t, \ldots, u_i^t, \ldots, u_N^t, vm_{cpu}]$. Here, u_i^t denotes the CPU utilization of the i-th PM, and vm_{cpu} represents the CPU request information of the VM to be placed.

2) Action selection: Action a_t refers to placing the selected VM onto an appropriate PM. To balance exploration and exploitation, the upper confidence bound (UCB) strategy is adopted to improve action selection. For a given environmental state s_t, the optimal action is selected according to Eq. (9):

$$a_t^* = \underset{a \in A}{argmax}\, UCB(s_t,a) \quad (9)$$

where $\mathrm{UCB(s_t,a)}$ is defined as:

$$UCB(s_t,a) = Q_\theta(s_t,a) + c \cdot \sqrt{\ln N(s_t)/N(s_t,a)} \quad (10)$$

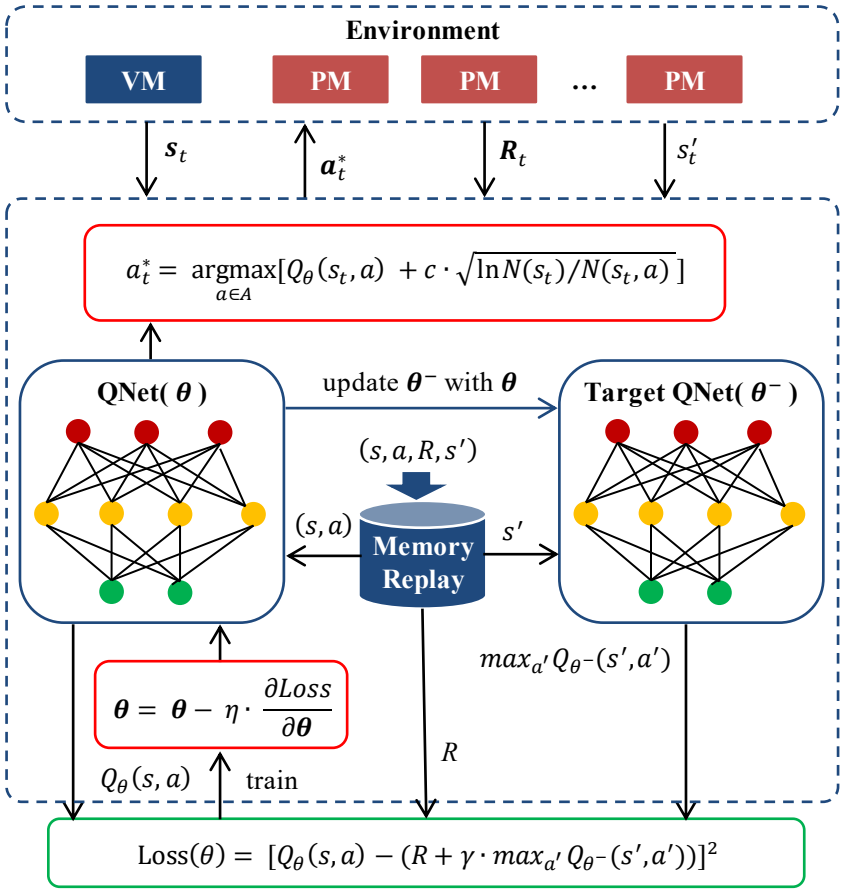


Fig. 3. VM placement framework based on enhanced deep Q-Network.

where $Q_\theta(s_t, a)$ is the value estimation of action a under state s_t from the online Q-network; c is the exploration coefficient; $N(s_t)$ represents the total number of times the state s_t has been visited; and $N(s_t, a)$ is the number of times the action a has been selected under state s_t. In this way, more diverse VM placement actions can be explored in the early stage. As experience accumulates, the focus gradually shifts to exploiting the optimal placement strategies that have been identified.

3) Reward: Based on the multi-objective optimization problem in the system model, the reward function is expressed as a function of energy consumption, SLAV rate, and migration count, which is defined as follows:

$$R_t = -(\alpha \cdot \frac{EC - EC_{min}}{EC_{max} - EC_{min}} + \beta \cdot \frac{SLAV - SLAV_{min}}{SLAV_{max} - SLAV_{min}} + \gamma \cdot \frac{M - M_{min}}{M_{max} - M_{min}}) \quad (11)$$

where $\alpha, \beta, \gamma \in [0, 1]$ are weight coefficients that satisfy $\alpha + \beta + \gamma = 1$; EC denotes total energy consumption, EC_{max} and EC_{min} represent the theoretical maximum and minimum energy consumption respectively; $SLAV_{max}$ and $SLAV_{min}$ are the theoretical maximum and minimum SLAV respectively; M indicates the migration count, M_{max} and M_{min} are the theoretical maximum and minimum migration counts respectively.

4) Loss Function: The online Q-network uses a mean squared error (MSE) loss function to calculate parameter gradients, which is defined by the following formula:

$$Loss(\theta) = E[(Q_\theta(s, a) - (R + \gamma \cdot max_{a'} Q_{\theta^-}(s', a')))^2] \quad (12)$$

where $Q_\theta(s, a)$ is the value estimation of the online Q-network for performing action a under state s; R denotes the obtained reward; $\gamma \in [0,1)$ is the discount factor; $Q_{\theta^-}(s', a')$ represents the value estimation of the target Q-network for action a′ under the next state s′; and $max_{a'} Q_{\theta^-}(s', a')$ refers to selecting the optimal action for the next state through the target Q-network.

5) Adaptive learning rate: An adaptive learning rate (ALR) is used for the training of the online Q-network, which is defined as follows:

$$\eta = \frac{1}{N(s,a)+1} \quad (13)$$

where $N(s, a)$ denotes the number of times action a has been selected under state s.

When the number of visits to a state-action pair is extremely small, $N(s, a) \approx 0$, and the learning rate $\eta \approx 1$ accordingly. A larger step size allows the parameters of the online Q-network to update rapidly, accelerating the learning process for value estimation of new decisions. As the same state-action pair is visited repeatedly, $N(s, a)$ increases, and the

learning rate η gradually approaches 0. A smaller step size prevents parameter oscillations caused by excessive updates, enabling the Q-value estimation to converge to a stable value and ensuring the accurate exploitation of known high-quality strategies.

B. VM Consolidation Algorithm

The VM consolidation algorithm is shown in Algorithm 3. The consolidation process consists of three parts: workload detection, VM detection, and VM placement. For each time window, the DetectWorkload function is executed first to detect the state of each PM. For overloaded PM, the SelectVMS function is executed to select the VMs that need to be migrated out. For underloaded PM, all VMs are migrated out. The VM placement phase is the core decision-making part of the algorithm, integrating deep reinforcement learning and the exploration-exploitation balance strategy.

The action selection adopts an improved variant of the ε-greedy strategy. On the basis of Q-network valuation, it introduces a confidence interval adjustment term composed of the exploration coefficient c, state-action counters N(s, a) and N(s). This design not only enables the exploitation of known high-quality placement schemes through the Q-network, but also encourages the exploration of low-frequency visited actions via $\sqrt{\ln N(s_t)/N(s_t,a)}$, thus preventing the algorithm from falling into local optimal solutions.

Algorithm 3. Virtual Machine Consolidation

Initialize:
the size of time window
the physical machines $H = \{H_1, H_2, \dots, H_N\}$
online net QNet(θ) with random weights θ
target net TargetQNet(θ^-) with weights $\theta^- \leftarrow \theta$
experience replay buffer D with capacity C
exploration coefficient c
discount factor γ

For each time step t **Do**
  Overload,Underload,Normal ← **DetectWorkload**(H)
  MigrationList ← **SelectVMS**(Overload)
  MigrationList ←*MigrationList*+VMs of underload PMs
  For each virtual machine V in *MigrationList* **Do**
    s_t = getCurrentState(*Normal*, V)
    $a_t^* = \underset{a \in A}{\operatorname{argmax}} [Q_\theta(s_t, a) + c \cdot \sqrt{\ln N(s_t)/N(s_t,a)}]$
    Place V on the host determined by the action a_t^*
    $R_t = -(\alpha \frac{EC - EC_{min}}{EC_{max} - EC_{min}} + \beta \frac{SLAV - SLAV_{min}}{SLAV_{max} - SLAV_{min}} + \gamma \frac{M - M_{min}}{M_{max} - M_{min}})$
    D ← D + (s, a, R, s')
  End For
  If D.size()>=batchsize AND (t%traininterval)==0 **Then**
    Batch = D.sample(batch size)
    For each sample (s, a, R, s') in Batch **Do**
      $\text{Loss}(\theta) = [Q_\theta(s,a) - (R + \gamma \cdot max_{a'} Q_{\theta^-}(s', a'))]^2$
      $\eta = \frac{1}{N(s,a)+1}$
      $\boldsymbol{\theta} = \boldsymbol{\theta} - \eta \cdot \frac{\partial Loss}{\partial \boldsymbol{\theta}}$
    End For
  End If
  If (t % target update interval) ==0 **Then**
    $\theta^- \leftarrow \theta$
  End If
End For

The experience replay mechanism caches historical interaction data and performs random sampling for training when the batch condition is met. This breaks the temporal correlation between samples and improves the stability of Q-network training. The loss function adopts the squared temporal difference error, taking the deviation between the current Q-value and the target Q-value as the optimization objective. The learning rate η is linked to N(s, a), enabling state-action pairs that are less frequently visited to obtain a larger parameter update magnitude, thus accelerating the learning of sparse sample regions.

The periodic update of the target network further prevents oscillations in Q-value estimation. By copying the online network parameters θ to the target network parameters θ^- at fixed intervals, the calculation benchmark for target Q-values remains stable within a certain period, providing a reliable reference for the gradient descent of the online network.

Overall, through the closed-loop mechanism of perception-decision-learning, this algorithm achieves dynamic optimization of VM consolidation. Workload detection accurately identifies resource bottlenecks; VM selection focuses on key adjustment targets; VM placement integrates reinforcement learning to realize intelligent decision-making; and the design of experience replay and target network ensures the convergence and robustness of the algorithm in dynamic cloud environments. Ultimately, it achieves the comprehensive goals of improving resource utilization, reducing data center energy consumption, and ensuring service quality.

VII. Experimental Evaluation

A. Experimental Setting

To verify the effectiveness of the adaptive VM consolidation strategy based on Autoformer and Enhanced Double Q-Network (AEDQN-VMC) proposed in this paper, the experimental simulations were conducted by building a test environment on the CloudSim 4.0 simulation platform [30]. As a mainstream simulation tool in the cloud computing field, this platform supports the simulation of resource scheduling for PMs and VMs, as well as the quantitative calculation of key metrics such as energy consumption and SLA violation rate, and has been widely adopted in relevant research. To cover the load characteristics under different business scenarios and

verify the generalization ability of the strategy, three types of real cloud load datasets were used in the experiments: the PlanetLab dataset [31], the Google Cluster Trace dataset [32], and the Alibaba dataset [33].

1) Configuration of PMs and VMs: To ensure consistency with the experimental conditions of existing mainstream research, the hardware parameters of PMs and VMs are all set with reference to the experimental parameters in [5]. The specific configurations are as follows:

a) PMs: Two types of heterogeneous hosts (HP ProLiant G4 and HP ProLiant G5) are adopted. Each host is configured with 2 CPU cores and a unified memory capacity of 4GB. The core difference lies in processor performance, with specific parameters shown in Table I. A total of 100 physical machines are deployed in the data center in the experiment, and the quantity ratio of the two types of hosts is 1:1, so as to simulate the heterogeneous hardware environment of real data centers.

TABLE I. PM CONFIGURATION

PM Type	Processor (MIPS)	Num. of Cores	Memory (GB)
HP ProLiant G4	1860	2	4
HP ProLiant G5	2660	2	4

b) VMs: Four types of VM instances are designed to cover scenarios from lightweight to high-performance. The MIPS (Million Instructions per Second) of processor ranges from 500 to 2500, and the memory capacity matches the CPU requirements, with specific parameters shown in Table II. There are 200 VMs in total in the experiment, with 50 VMs of each type.

TABLE II. VM CONFIGURATION

VM Type	Processor(MIPS)	Memory(GB)
Micro	500	0.85
Small	1000	1.7
Extra Large	2000	3.75
High-CPU Medium	2500	0.85

2) Parameters of the energy consumption model: The energy consumption calculation of PMs refers to the SPECpower benchmark test data [34]. As an industry standard for data center energy consumption evaluation, this benchmark provides measured power consumption values under different CPU utilization rates. In the experiment, the power consumption data of the two types of hosts within the CPU utilization range of 0%–100% is shown in Table III. For intermediate utilization rates not covered in the table (e.g., 15%, 25%), the linear interpolation method is used to calculate continuous power consumption values, ensuring the accuracy of the energy consumption model.

3) Key parameter settings: Autoformer Load Prediction Parameters: The length of the historical time window is 24 time steps; the number of encoder/decoder layers is 2 each; the number of heads in the autocorrelation block of each layer is 4; and the hidden layer dimension is 64. The batch size is 128, the number of epochs is 100, the Adam optimizer is used, the initial learning rate is 0.01, and the loss function is mean squared error (MSE). The overload threshold of physical machines is 0.9, and the underload threshold is 0.2.

Enhanced Double Q-Network Parameters: A fully connected neural network is adopted as the network model for the double Q-network, and the online Q-network and target Q-network share an identical network structure. This fully connected neural network contains 3 hidden layers, with the number of nodes in each hidden layer being 256, 128, and 64 respectively. The exploration coefficient c in (10) is 0.5; α, β, γ in (11) are each 1/3; and the discount factor γ in (12) is 0.9. The training interval is 10 time steps, and the target network update interval is 100 time steps.

TABLE III. POWER CONSUMPTION AT DIFFERENT CPU UTIL. (WATTS)

CPU Util.	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
HP G4	86	89.4	92.6	96	99.5	102	106	108	112	114	117
HP G4	93.7	97	101	105	110	116	121	125	129	133	135

4) Comparison benchmarks and evaluation metrics: To verify the superiority of the AEDQN-VMC strategy, three VM consolidation strategies were selected as comparison benchmarks in the experiment:

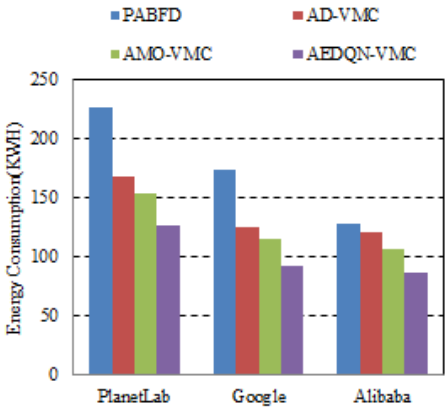
- PABFD (Power-Aware Best Fit Decreasing): An energy-aware strategy built into CloudSim [35], which selects the host with the minimum energy consumption increment after VM placement.
- ADVMC (Adaptive DRL-based VM Consolidation): A VM consolidation strategy proposed by [5], which is based on LSTM prediction and DQN placement.
- AMOVMC (Adaptive Multi-Objective Virtual Machine Consolidation): An VM consolidation strategy proposed by [4], which uses a neural network to predict future resource utilization, and adopts a multi-objective heuristic adaptive VM placement algorithm to select the optimal target host.

Two core evaluation metrics were used in the experiment: total energy consumption and SLAV rate. The calculation of total energy consumption refers to Eq. (2), and the calculation of SLAV rate refers to Eq. (3)–(5). The CPU performance loss caused by migration is set to 10% by default.

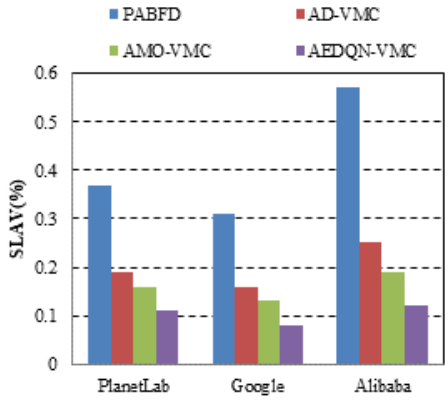
B. Performance Evaluation

To verify the effectiveness and advancement of the VM consolidation method AEDQN-VMC proposed in this paper, we conducted a comparative analysis between the proposed method and the benchmark methods. The comparison results are shown in Fig. 4. It can be seen that the energy consumption of the method proposed in this paper is significantly lower than that of the comparison methods on the three datasets (PlanetLab, Google, and Alibaba). Specifically, compared with the PABFD method, the energy consumption is reduced by 44.5%, 46.5%, and 32.6% respectively; compared with the ADVMC method, the energy consumption is reduced by

25.2%, 25.7%, and 28.3% respectively; and compared with the AMOVMC method, the energy consumption is reduced by 18%, 19.7%, and 18.8% respectively.



(a) Energy



(b) SLAV

Fig. 4. Comparison of performance among consolidation methods.

The performance improvement in energy consumption stems from the following aspects. The Autoformer-based workload detection module can accurately predict the load trend of PMs, identify overloaded and underloaded states in advance, and avoid redundant energy consumption caused by unbalanced resource allocation. The VM placement module achieves efficient resource utilization through the enhanced double Q-network. On the premise of meeting the resource requirements of VMs, it minimizes the number of active PMs, thereby reducing overall energy consumption. The timely shutdown strategy for underloaded hosts further reduces idle energy consumption and improves energy utilization efficiency.

The SLAV of the method proposed in this paper remains the lowest across all three datasets. Specifically, on the Google dataset, its SLAV is 0.08%, a reduction of 50% compared to the ADVMC method, a reduction of 38.5% compared to the AMOVMC method, and a reduction of 74.2% compared to the PABFD method. The accurate load prediction of the Autoformer model can avoid the risk of PM overload in advance and reduce service interruptions caused by resource saturation. The VM selection module screens migration targets using the Pearson correlation coefficient, prioritizing the migration of VMs that have less impact on host load. Meanwhile, by integrating migration time optimization, it reduces performance loss during the migration process, thereby lowering the overall SLAV.

C. Ablation Experiments

1) Impact of PM detection on consolidation: To verify the impact of the Autoformer-based host state detection (Autoformer-HSD) proposed in this paper on VM consolidation performance, we kept the VM selection algorithm and VM placement algorithm unchanged, and replaced the host state detection algorithm proposed in this paper with other host state detection algorithms. In this way, we obtained the VM consolidation performance based on different host state detection algorithms. The benchmark methods used to verify Autoformer-HSD are as follows:

- Current state-based host state detection without prediction (CS-HSD)
- Local regression-based host state detection (LR-HSD)
- LSTM-based host state detection (LSTM-HSD)

The performance comparison results between Autoformer-HSD and other methods are shown in Fig. 5. As can be seen from Fig. 5, Autoformer-HSD is significantly superior to other comparison methods in terms of the two core metrics—energy consumption and SLAV, verifying its role in improving VM consolidation performance.

a) Energy consumption: Compared with CS-HSD, the energy consumption of Autoformer-HSD is reduced by 30.8% (PlanetLab), 34.5% (Google), and 26.4% (Alibaba) across the three datasets. This is because CS-HSD only relies on the current load state to make decisions and cannot identify overload risks in advance. It often triggers migration only after the host is already overloaded, leading to redundant resource allocation and additional energy consumption from frequent migrations. In contrast, Autoformer-HSD can adjust resource allocation in advance through accurate prediction of future loads, reducing ineffective energy consumption at the source.

Compared with LR-HSD, the energy consumption of Autoformer-HSD is reduced by 25.6% (PlanetLab), 19.0% (Google), and 22.5% (Alibaba). LR-HSD can only capture local linear trends and has limited ability to model complex periodic fluctuations and long-term trends in cloud loads, resulting in high prediction errors that further affect the accuracy of host state judgment. However, the autocorrelation mechanism of Autoformer can effectively decompose the periodic and trend components of the load, improving prediction accuracy and reducing energy waste caused by misjudgment.

Compared with LSTM-HSD, the energy consumption of Autoformer-HSD is still reduced by 5.3% (PlanetLab), 7.1% (Google), and 7.7% (Alibaba). This benefit comes from the global perception mechanism of the Autoformer decoder, enabling it to more accurately capture the long-term dependencies of the load. This optimizes the timeliness and accuracy of host state detection, further reducing energy consumption.

b) SLAV: The SLAV of Autoformer-HSD is the lowest across all datasets. Compared with CS-HSD, its SLAV is reduced by 57.7% (PlanetLab), 61.9% (Google), and 36.8% (Alibaba). This is because CS-HSD lacks prediction capability and cannot avoid overload in advance, causing hosts to frequently be in a state of resource saturation and triggering SLA violations. In contrast, Autoformer-HSD can complete VM migration before the arrival of load peaks by predicting overload risks in advance, significantly reducing service interruption time.

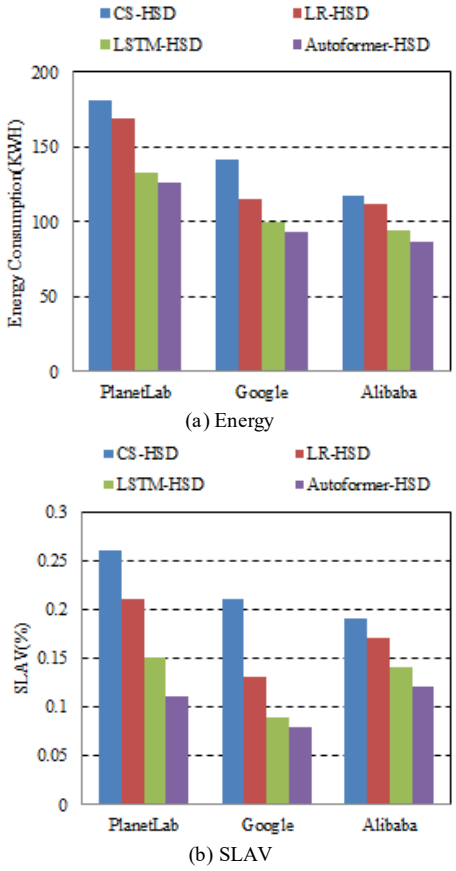


Fig. 5. Performance comparison between Autoformer-HSD and other host detection methods.

Compared with LR-HSD and LSTM-HSD, the SLAV advantage of Autoformer-HSD is slightly narrowed but still maintains a leading position. For example, on the Google dataset, its SLAV is 38.5% lower than that of LR-HSD and 11.1% lower than that of LSTM-HSD. This stems from Autoformer's accurate modeling of load fluctuations, which enables more precise judgment on whether a host will enter an overloaded state, reducing the increase in SLAV caused by prediction errors.

This ablation experiment shows that Autoformer-HSD can more precisely identify the overloaded and underloaded states of hosts by improving the accuracy of load prediction and the ability to model complex time-series features. It provides a reliable decision-making basis for subsequent VM selection and placement, ultimately reducing energy consumption while effectively ensuring service quality. This verifies its core role in the VM consolidation system.

2) Impact of VM selection on consolidation: To evaluate the impact of the PCM-VMS proposed in this paper on the overall performance of VM consolidation, we kept all other components of the proposed VM consolidation algorithm unchanged, while replacing PCM-VMS with other benchmark VM selection algorithms. In this way, we obtained performance data of VM consolidation based on different VM selection algorithms. The other benchmark VM selection algorithms used to verify the performance of PCM-VMS are as follows:

- Minimum migration time-based VM selection (MMT-VMS)
- Maximum correlation-based VM selection (MC-VMS)
- Influence coefficient-based VM selection (IC-VMS)

The performance comparison results between PCM-VMS and other VM selection algorithms are shown in Fig. 6. It can be indicated that the proposed PCM-VMS performs the best in terms of both energy consumption and SLAV metrics, verifying its role in improving VM consolidation performance.

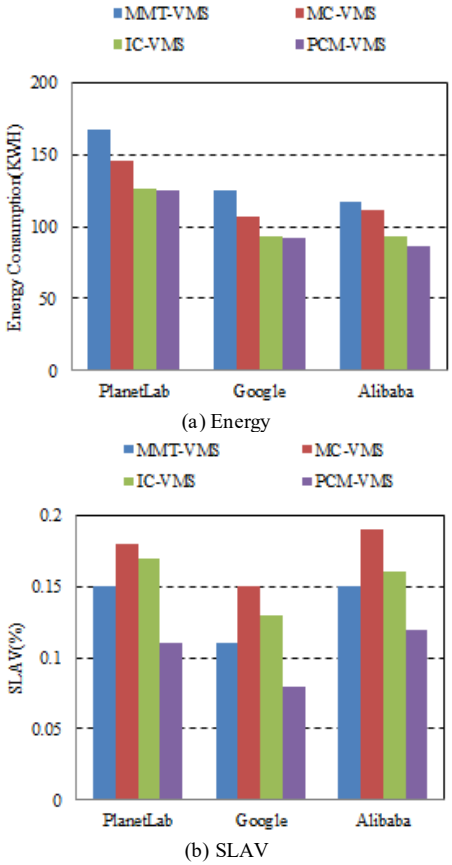


Fig. 6. Performance comparison between PCM-VMS and other VM selection algorithms.

a) Energy consumption: Compared with MMT-VMS which only considers migration time, the energy consumption of PCM-VMS is reduced by 25.0% (PlanetLab), 26.1%

(Google), and 26.4% (Alibaba) across the three datasets. Although MMT-VMS can reduce the time cost of a single migration, it fails to consider the resource correlation between VMs and hosts. This may lead to the migration of VMs that have little impact on host load, resulting in overloaded hosts not being effectively relieved of their load, and requiring multiple migrations to achieve load balancing, which instead increases overall energy consumption. In contrast, PCM-VMS balances correlation and migration cost, which enable overloaded hosts to return to a normal state through fewer migration operations, thereby reducing energy consumption.

Compared with MC-VMS which only focuses on correlation, the energy consumption of PCM-VMS is reduced by 14.1% (PlanetLab), 13.1% (Google), and 22.5% (Alibaba). Although MC-VMS can quickly alleviate host overload by migrating VMs with high correlation, it fails to consider migration time. This may lead to system performance loss due to excessive network resource occupation during migration, indirectly increasing energy consumption. In contrast, PCM-VMS penalizes the migration of VMs with large memory, reducing migration operations that incur high resource overhead and further optimizing energy consumption.

Compared with IC-VMS based on the influence coefficient, the energy consumption of PCM-VMS is reduced by 7.7% on the Alibaba dataset, and its performance is comparable to that of IC-VMS on the other datasets. This is because the influence coefficient of IC-VMS does not incorporate migration cost, whereas PCM-VMS precisely controls migration cost through a denominator term, making its decisions more aligned with the core goals of efficient load reduction and low overhead.

b) SLAV: The SLAV of PCM-VMS is the lowest across all datasets. Compared with MMT-VMS, its SLAV is reduced by 26.7% (PlanetLab), 27.3% (Google), and 20.0% (Alibaba), respectively. Since MMT-VMS ignores the correlation between VMs and hosts, it may cause host load rebound after migration, increasing SLAV; in contrast, the high-correlation VMs migrated by PCM-VMS can effectively reduce host load fluctuations and lower the risk of overload.

Compared with MC-VMS, the SLAV of PCM-VMS is reduced by 38.9% (PlanetLab), 46.7% (Google), and 36.8% (Alibaba). MC-VMS prioritizes migrating high-correlation VMs but may select VMs with large memory, leading to a decline in migration performance. In contrast, PCM-VMS restricts the migration of large-memory VMs through a denominator term, reducing performance loss during the migration process and thereby lowering the overall SLAV.

Compared with IC-VMS, the SLAV of PCM-VMS is reduced by 35.3% (PlanetLab), 38.5% (Google), and 25.0% (Alibaba). This benefit comes from PCM-VMS's refined modeling of correlation and migration cost, which ensures that migration operations can effectively alleviate host overload and control migration time. The collaborative optimization of these two aspects significantly reduces SLAV.

In summary, this ablation experiment verifies the effectiveness of PCM-VMS. By integrating the Pearson correlation coefficient and migration time, a more optimized influence factor is constructed. While reducing the number of migrations and lowering energy consumption, it effectively controls SLAV, providing an efficient decision-making basis for VM consolidation.

3) Impact of VM placement on consolidation: To evaluate the performance of the VM placement algorithm based on enhanced dual Q network (EDN-VMP) proposed in this paper, we kept all other components of the proposed VM consolidation algorithm unchanged, while replacing EDN-VMP with other benchmark VM placement algorithms. In this way, we obtained performance data of VM consolidation based on different VM placement algorithms. The other benchmark VM placement algorithms used to verify the performance of EDN-VMP are as follows:

- Minimum energy increment based VM placement (MEI-VMP)
- Q-Learning-based VM placement (QL-VMP)
- Dual-Q-network-based VM placement (DN-VMP)

The comparison results of VM consolidation performance between the EDN-VMP and other VM placement algorithms are shown in Fig. 7. It can be seen that the EDN-VMP is significantly superior to other benchmark algorithms in both energy consumption and service quality, fully verifying its role in improving VM consolidation performance.

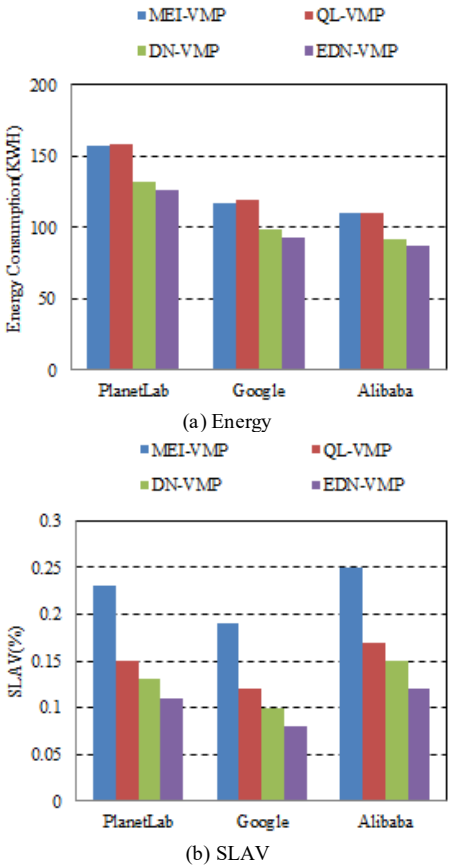


Fig. 7. Performance comparison between EDN-VMP and other VM placement algorithms.

a) Energy consumption: Compared with MEI-VMP, the energy consumption of EDN-VMP is reduced by 20.0% (PlanetLab), 21.1% (Google), and 21.1% (Alibaba) respectively. MEI-VMP adopts a greedy strategy and only takes the current energy consumption increment as the decision-making basis. It fails to consider long-term load trends and resource balance, which may lead to frequent migrations or host overload in the later stage due to short-term optimization, thereby increasing overall energy consumption. In contrast, EDN-VMP leverages the global decision-making capability of reinforcement learning and combines the UCB strategy to explore better placement schemes. While reducing current energy consumption, it avoids long-term resource imbalance, thus achieving more stable energy consumption optimization.

Compared with QL-VMP, the energy consumption of EDN-VMP is reduced by 20.6% (PlanetLab), 22.3% (Google), and 21.7% (Alibaba) respectively. QL-VMP has limited ability to model high-dimensional state spaces and struggles to accurately capture the correlation between placement decisions and long-term energy consumption, resulting in low decision-making accuracy. In contrast, EDN-VMP extracts complex state features through deep neural networks and combines a dual Q-network structure to avoid Q-value overestimation, significantly enhancing the global optimization capability of its decisions.

Compared with DN-VMP, the energy consumption of EDN-VMP is still reduced by 5.1% (PlanetLab), 6.1% (Google), and 5.8% (Alibaba). Although DN-VMP solves the overestimation problem, it lacks the key improvements of EDN-VMP: (1) It does not introduce the UCB strategy, resulting in insufficient exploration of high-quality placement schemes with low access frequency and a tendency to fall into local optimality; (2) It does not adopt an adaptive learning rate, leading to low efficiency in parameter updates for sparse scenarios. Through these two improvements, EDN-VMP further enhances the accuracy of resource allocation and reduces redundant energy consumption.

b) SLAV: Compared with MEI-VMP, the SLAV of EDN-VMP is reduced by 52.2% (PlanetLab), 57.9% (Google), and 52.0% (Alibaba), respectively. MEI-VMP, due to its over-pursuit of energy consumption optimization, may deploy VMs to hosts that are nearly saturated, leading to frequent host overload. Additionally, it lacks global migration planning, which may trigger unnecessary migration operations and result in decreased migration performance. In contrast, the reward function of EDN-VMP constrains both energy consumption and SLAV simultaneously. By balancing multi-objectives through reinforcement learning, it effectively reduces the risk of overload and migration losses.

Compared with QL-VMP, the SLAV of EDN-VMP is reduced by 26.7% (PlanetLab), 33.3% (Google), and 29.4% (Alibaba) respectively. QL-VMP's coarse state modeling leads to low matching accuracy between placement decisions and VM resource requirements, easily causing overload. However, EDN-VMP accurately captures state features via deep networks and explores better matching schemes by combining the UCB strategy, significantly reducing the SLA violation rate.

Compared with DN-VMP, the SLAV of EDN-VMP is reduced by 15.4% (PlanetLab), 20.0% (Google), and 20.0% (Alibaba) respectively. The performance gap arises because DN-VMP lacks an adaptive learning rate. For newly emerged state-action pairs, the fixed parameter update step size of DN-VMP makes it difficult to quickly learn optimal decisions, leading to a short-term increase in SLAV. In contrast, EDN-VMP assigns a larger update magnitude to sparse samples through its adaptive learning rate, accelerating adaptation to new scenarios and thereby reducing the violation rate.

In summary, this ablation experiment fully verifies the effectiveness of EDN-VMP. By introducing the UCB strategy to balance exploration and exploitation, adopting an adaptive learning rate to optimize parameter updates, and integrating the global decision-making capability of the dual Q-network, EDN-VMP significantly improves the stability of service quality while reducing energy consumption.

4) Impact of UCB and ALR on Dual Q-Network: To verify the impact of the UCB and ALR on the performance of the Dual Q-Network, we further conducted ablation experiments on the enhanced dual Q-Network. Based on the basic dual Q-Network, we integrated the UCB and adaptive learning rate, respectively. The several neural networks used for comparison are as follows:

- General dual Q-Network (DQN)
- Dual Q-Network with UCB (DQN+UCB)
- Dual Q-Network with ALR (DQN+ALR)
- Dual Q-Network with UCB and ALR (DQN+UCB+ ALR)

The impact of the UCB and ALR on the performance of the Dual Q-Network is shown in Table IV. The experimental results indicate that the introduction of UCB and ALR plays a significant role in improving the performance of the Dual Q-Network, and the synergistic effect of the two further optimizes the core metrics of VM consolidation.

TABLE IV. IMPACT OF UCB AND ALR ON PERFORMANCE OF DUAL Q-NETWORK

	Energy Consumption (KWH)			SLAV (%)		
	PlanetLab	Google	Alibaba	PlanetLab	Google	Alibaba
DQN	132.4	98.7	91.8	0.13	0.1	0.15
DQN+UCB	129.8	95.8	89.4	0.12	0.09	0.13
DQN+ALR	128.6	94.6	88.7	0.12	0.09	0.13
DQN+UCB+ALR	125.6	92.65	86.52	0.11	0.08	0.12

In terms of energy consumption, compared with the general dual Q-Network (DQN), the DQN+UCB achieves an energy consumption reduction of 1.96% (PlanetLab), 2.94% (Google), and 2.61% (Alibaba) across the three datasets. This benefit stems from the UCB strategy's ability to balance exploration and exploitation, which reduces resource allocation imbalance caused by local optimality and makes VM placement decisions more aligned with the goal of global energy consumption optimization. The DQN+ALR shows a more significant reduction in energy consumption. Its advantage lies in its fast learning capability for sparse state-action pairs, which accelerates the convergence of high-quality placement strategies and reduces redundant energy consumption.

The DQN+UCB+ALR performs the best, with energy consumption reduced by 5.14% (PlanetLab), 6.13% (Google), and 5.75% compared with DQN, reflecting the synergistic effect of the two mechanisms. UCB ensures sufficient exploration of the decision space, while ALR improves learning efficiency, enabling the algorithm to quickly find the balance between energy consumption and service quality in dynamic cloud environments.

In terms of SLAV, the SLAV of DQN+UCB is reduced by 7.69% (PlanetLab), 10.00% (Google), and 13.33% (Alibaba) compared with DQN. This indicates that the UCB strategy reduces the risk of host overload caused by decision limitations by exploring more potential high-quality placement schemes. The SLAV performance of DQN+ALR is comparable to that of DQN+UCB. This verifies the ALR's ability to quickly adapt to new scenarios, which can effectively reduce short-term service interruptions.

The SLAV advantage of DQN+UCB+ALR is further expanded. Compared with DQN, its SLAV is reduced by 15.38% (PlanetLab), 20.00% (Google), and 20.00% (Alibaba). The reason lies in that the UCB strategy reduces the proportion of suboptimal decisions caused by insufficient exploration, while the ALR accelerates the learning of high-risk state-action pairs. The combined effect of the two enables the algorithm to significantly reduce the SLAV rate.

In summary, the introduction of the UCB and ALR optimizes the dual Q-Network from two dimensions: the breadth of decision exploration and the depth of learning efficiency. Their combination achieves a "1+1>2" optimization effect through a synergistic effect, providing more accurate and efficient intelligent support for VM placement decisions. Ultimately, it significantly improves the stability of service quality while reducing energy consumption.

VIII. Conclusion and Future Work

To address the issues of energy optimization and service quality management in VM consolidation for cloud data centers, this paper proposes an adaptive consolidation strategy based on Autoformer and enhanced dual Q-Network. Through collaborative innovations in three stages: host state detection, VM selection, and VM placement, an effective balance between energy consumption and service quality in dynamic cloud environments is achieved.

In the host state detection stage, a prediction method based on Autoformer is adopted. This method performs multi-scale decomposition of load time series through an autocorrelation mechanism, explicitly separating trend components from periodic components. This significantly improves both load prediction accuracy and the accuracy of host state classification. In the VM selection stage, a VM migration impact factor that integrates the Pearson correlation coefficient and migration time is proposed. This factor balances resource correlation and migration overhead, effectively avoiding performance degradation caused by excessively high migration costs or insufficient load mitigation. In the VM placement stage, an enhanced dual Q-Network algorithm is designed. It introduces the UCB strategy to optimize the exploration-exploitation trade-off and adopts an ALR mechanism to accelerate model convergence, thereby achieving intelligent decision-making under multi-objective trade-offs.

Experimental results show that the strategy proposed in this paper significantly outperforms the comparative benchmark methods on three real load datasets, and can effectively reduce cloud data center energy consumption and SLAV rates. Ablation experiments further verify the effectiveness of each innovative module and their contributions to the overall performance.

Despite the good performance of the AEDQN-VMC strategy, there are still some limitations that can be further explored in future research:

- Adaptation to multi-resource dimensions and heterogeneous workloads: The current study mainly focuses on CPU and memory resources. In the future, it can be extended to multi-dimensional resource constraints such as network I/O and disk bandwidth, and the adaptability to heterogeneous workloads can be enhanced.
- Model lightweighting and online learning: The Autoformer and Dual Q-Network models have a large number of parameters, resulting in high training and inference overhead. Future research can explore model compression, knowledge distillation, and online incremental learning mechanisms to improve the practicality and real-time performance of the algorithm in ultra-large-scale data center environments.
- Cross-layer collaboration and cooling energy consumption optimization: The current energy consumption model only considers server energy consumption. In the future, a cooling system energy consumption model can be introduced, and cross-layer collaborative optimization between computing resource scheduling and cooling system control can be explored to further improve overall energy efficiency.
- Extension to multi-cloud and edge environments: This study focuses on a single data center scenario. In the future, the framework can be extended to multi-cloud collaboration or edge-cloud collaboration environments, and cross-domain resource scheduling and fault-tolerance mechanisms can be studied to enhance the generalization ability and application scope of the strategy.

ACKNOWLEDGMENT

This work was supported in part by Heilongjiang Provincial Natural Science Foundation Funded Project under Grant PL2024F003, and in part by Heilongjiang Provincial General Project of College Students' Innovation and Entrepreneurship Training Program under Grant S202510240008.

REFERENCES

[1] S. S. Gill, R. Buyya, "A taxonomy and future directions for sustainable cloud computing: 360 degree view." ACM Computing Surveys, vol. 51, no. 5, pp. 1-33, 2019.

[2] A. H. T. Dias, L. H. A. Correia, N. Malheiros, "A systematic literature review on virtual machine consolidation." ACM Computing Surveys, vol. 54, no. 8, pp. 1-38, 2021.

[3] J. Singh, N. K. Walia, "A comprehensive review of cloud computing virtual machine consolidation." IEEE Access, vol. 11, pp. 106190-106209, 2023.

[4] S. Goyal, L. K. Awasthi, "Adaptive multi-objective virtual machine consolidation for energy-efficient cloud data centers." Journal of Grid Computing, vol. 23, no. 2, article no. 21, 2025.

[5] J. Zeng, D. Ding, K. Kang, H. Xie, Q. Yin, "Adaptive drl-based virtual machine consolidation in energy-efficient cloud data center." IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 11, pp. 2991-3002, 2022.

[6] B. Feng, Z. Ding, "Application-oriented cloud workload prediction: A survey and new perspectives." Tsinghua Science and Technology, vol. 30, no. 1, pp. 34-54, 2025.

[7] F. Zhao, W. Lin, S. Lin, H. Zhong, K. Li, "Tfegru: Time-frequency enhanced gated recurrent unit with attention for cloud workload prediction." IEEE Transactions on Services Computing, vol. 18, no. 1, pp. 467-478, 2025.

[8] M. Zakarya, L. Gillam, M. R. C. Qazani, A. A. Khan, K. Salah, et al., "Backfillme: An energy and performance efficient virtual machine scheduler for iaas datacenters." IEEE Transactions on Services Computing, vol. 18, no. 2, pp. 660-672, 2025.

[9] N. Ma, A. Tang, Z. Xiong, F. Jiang, "A deep multi-agent reinforcement learning approach for the micro-service migration problem with affinity in the cloud." Expert Systems with Applications, vol. 273, article no. 126856, 2025.

[10] D. Zhao, J. t. Zhou, K. Li, "Cfws: Drl-based framework for energy cost and carbon footprint optimization in cloud data centers." IEEE Transactions on Sustainable Computing, vol. 10, no. 1, pp. 95-107, 2025.

[11] G. Long, S. Wang, C. Lv, "Qos-aware resource management in cloud computing based on fuzzy meta-heuristic method." Cluster Computing, vol. 28, no. 4, article no. 276, 2025.

[12] J. Li, Y. Deng, R. Wang, Y. Zhou, H. Feng, et al., "Btvmp: A burst-aware and thermal-efficient virtual machine placement approach for cloud data centers." IEEE Transactions on Services Computing, vol. 17, no. 5, pp. 2080-2094, 2024.

[13] X. Zhang, T. Wu, M. Chen, T. Wei, J. Zhou, et al., "Energy-aware virtual machine allocation for cloud with resource reservation." Journal of Systems and Software, vol. 147, pp. 147-161, 2019.

[14] J. Li, Y. Deng, Z. Zhong, Z. Wu, S. Pang, et al., "An energy-aware virtual machine scheduling approach for cloud data centers." IEEE Transactions on Sustainable Computing, vol. 10, no. 5, pp. 891-907, 2025.

[15] Y. Chen, Z. Zhang, Y. Deng, G. Min, L. Cui, "A combined trend virtual machine consolidation strategy for cloud data centers." IEEE Transactions on Computers, vol. 73, no. 9, pp. 2150-2164, 2024.

[16] W. Yao, Z. Wang, Y. Hou, X. Zhu, X. Li, et al., "An energy-efficient load balance strategy based on virtual machine consolidation in cloud environment." Future Generation Computer Systems, vol. 146, pp. 222-233, 2023.

[17] R. Shaw, E. Howley, E. Barrett, "Applying reinforcement learning towards automating energy efficient virtual machine consolidation in cloud data centers." Information Systems, vol. 107, article no. 101722, 2022.

[18] P. Qin, C. Guo, "Construction of virtualization adaptive algorithm for virtual machine environment," in Proc. of 2025 3rd International Conference on Integrated Circuits and Communication Systems (ICICACS), 2025, pp. 1-6.

[19] S. Vila, F. Guirado, J. L. Lérida, "Cloud computing virtual machine consolidation based on stock trading forecast techniques." Future Generation Computer Systems, vol. 145, pp. 321-336, 2023.

[20] C. Guo, L. Li, M. Zukerman, "Q-learning-based workload consolidation for data centers with composable architecture." IEEE Transactions on Industrial Informatics, vol. 21, no. 3, pp. 2324-2333, 2025.

[21] N. Moocheet, B. Jaumard, P. Thibault, L. Eleftheriadis, "Minimum-energy virtual machine placement using embedded sensors and machine learning." Future Generation Computer Systems, vol. 161, pp. 85-94, 2024.

[22] S. Rahmani, V. Khajehvand, M. Torabian, "Burstiness-aware virtual machine placement in cloud computing systems." The Journal of Supercomputing, vol. 76, no. 1, pp. 362-387, 2020.

[23] Z. Tong, J. Wang, Y. Wang, B. Liu, Q. Li, "Energy and performance-efficient dynamic consolidate vms using deep-q neural network." IEEE Transactions on Industrial Informatics, vol. 19, no. 11, pp. 11030-11040, 2023.

[24] N. K. Sharma, S. Bojjagani, Y. C. A. P. Reddy, M. Vivekanandan, J. Srinivasan, et al., "A novel energy efficient multi-dimensional virtual machines allocation and migration at the cloud data center." IEEE Access, vol. 11, pp. 107480-107495, 2023.

[25] J.-Y. Luo, L. Chen, W.-K. Chen, J.-H. Yuan, Y.-H. Dai, "A cut-and-solve algorithm for virtual machine consolidation problem." Future Generation Computer Systems, vol. 154, pp. 359-372, 2024.

[26] A. Ignatov, I. Maslova, M. Posypkin, W. Yang, J. Wu, "A two-phase heuristic algorithm for power-aware offline scheduling in iaas clouds." Journal of Parallel and Distributed Computing, vol. 178, pp. 1-10, 2023.

[27] A. R. Hummaida, N. W. Paton, R. Sakellariou, "Scalable virtual machine migration using reinforcement learning." Journal of Grid Computing, vol. 20, pp. 1-26, 2022.

[28] K. Haghshenas, A. Pahlevan, M. Zapater, S. Mohammadi, D. Atienza, "Magnetic: Multi-agent machine learning-based approach for energy efficient dynamic consolidation in data centers." IEEE Transactions on Services Computing, vol. 15, no. 1, pp. 30-44, 2022.

[29] H. Wu, J. Xu, J. Wang, M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting." Advances in neural information processing systems, vol. 34, pp. 22419-22430, 2021.

[30] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. D. Rose, R. Buyya, "Cloudsim: A toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms." Software: Practice and Experience, vol. 41, no. 1, pp. 23-50, 2011.

[31] A. Beloglazov, "Planetlab workload traces," GitHub, 2011. https://github.com/beloglazov/planetlab-workload-traces.

[32] Google, "Google cluster workload traces (version 2011)," GitHub, 2011. https://github.com/google/cluster-data/blob/master/ClusterData2011_2.md.

[33] Alibaba, "Alibaba cluster trace (version 2018)," GitHub, 2018. https://github.com/alibaba/clusterdata/tree/master/cluster-trace-v2018.

[34] K. D. Lange, "Identifying shades of green: The specpower benchmarks." Computer, vol. 42, no. 3, pp. 95-97, 2009.

[35] A. Beloglazov, R. Buyya, "Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers." Concurrency and Computation: Practice & Experience, vol. 24, no. 13, pp. 1397–1420, 2012.

Handwriting Detectives Using Wavelet Siamese Technology to Verify Signature Fraud

Mohamed Nazir[1]*, Ali Maher[2], Mostafa Eltokhy[3]*, Ali M. El-Rifaie[4]*, Tarek Hosny[5], Hani M. K. Mahdi[6]*
Faculty of Engineering, Ain Shams University, Cairo, Egypt[1, 6]
Military Technical College, Cairo, Egypt[2]
Electronics Technology Department-Faculty of Technology and Education, Helwan University, Cairo, Egypt[3]
College of Engineering and Technology, American University of the Middle East, Egaila 54200, Kuwait[4]
Communication Engineering Department-Al-Safwa High Institute of Engineering, High Ministry of Education, Cairo, Egypt[5]

***Abstract*—This paper addresses the escalating challenge of signature forgery detection through an innovative hybrid verification system. We integrate Siamese Neural Networks with wavelet scattering transformations to precisely capture signature characteristics while accommodating inherent variations. Our principal contribution, the "common anchor methodology," identifies a singular representative signature per individual, substantially reducing computational demands on the CEDAR Dataset while maintaining verification integrity. Through meticulous optimization of wavelet scattering parameters, our system demonstrates markedly superior performance on the CEDAR benchmark while requiring considerably fewer model parameters than traditional CNN architectures. This research establishes noteworthy advancements in both accuracy and efficiency for practical signature verification implementations. The study evaluates the performance of a wavelet-Siamese network architecture for offline signature verification through a series of five experiments with varying parameter configurations. Key variables include the use of a common anchor, the J Factor, and the θ value. Results reveal that incorporating a common anchor consistently improves performance. Among all configurations, experiment 4 with a J Factor of 2 and a θ value of 16 yielded the most favorable results, achieving the lowest error rate of 20.823% and the highest ROC-AUC score of 0.8699, along with efficient convergence within 55 iterations. In contrast, the absence of a common anchor in Experiment 1 led to a notably higher error rate of 24.44% and lower model performance. These findings demonstrate the critical role of parameter tuning in enhancing the robustness and accuracy of signature verification systems based on Siamese networks. Despite the substantial computational savings, the system's best achieved error rate (20.82%) remains higher than several state-of-the-art and commercial signature verification solutions, many of which report error rates below 10%. This indicates an existing trade-off between efficiency and the highest attainable accuracy, which future work will aim to mitigate.**

***Keywords*—Biometric authentication; Siamese neural networks; scattering wavelets; common anchor selection; neutrosophic logic; signature verification**

I. Introduction

A signature, typically a person's name or identifying mark, serves as proof of document approval and plays a vital role in personal authentication. As a deliberately created biometric trait, handwritten signatures have become increasingly important with the rise of digital documents and transactions, where signature forgery poses a serious threat. Traditional manual verification methods are subjective, time-consuming, and prone to error. Therefore, there is a growing need for automated signature verification systems leveraging computer vision, machine learning, and biometric techniques. Standard CNN approaches particularly struggle with forgeries created by people who've practiced mimicking a target signature [1]. These sophisticated fakes often slip through undetected – a serious weakness in security-critical applications [2]. The computational demands also create barriers to widespread adoption, especially in resource-limited settings. Our work addresses these challenges using a novel approach. We employ Siamese neural networks enhanced with wavelet scattering transformations [3]. Instead of relying on rigid thresholds, the system learns adaptive similarity measures to accommodate natural variations. By examining signature features at multiple scales and angles, it significantly improves the detection of even well-crafted forgeries. Signature verification is categorized into dynamic (online) and static (offline) types. Dynamic methods capture signatures during writing on digital devices, while static methods analyze scanned images of completed signatures. Offline verification is particularly challenging due to the lack of dynamic information such as pen movement, speed, and pressure. Deep learning and machine learning techniques, especially Convolutional Neural Networks (CNNs), have been increasingly employed in recent studies [4], [5], [6], [7] to extract reliable and distinctive features, leading to significant improvements over traditional handcrafted approaches. Despite these advances, signature forgery remains a serious security challenge in the digital age [2]. Manual verification's subjectivity and inefficiency have driven interest in automated systems. Current CNN-based methods process signature images [8] and match them against stored samples using fixed thresholds. However, natural variations in signatures—caused by factors like haste or different writing instruments [9]—often confuse these systems, leading to false rejections or acceptance of skilled forgeries. Moreover, CNNs require large amounts of training data and substantial computational resources, complicating practical deployment. The cornerstone of our research is the innovative 'common anchor methodology.' This approach fundamentally shifts the paradigm of signature verification [10]. Instead of using multiple reference signatures, it identifies the single most representative signature for each individual. This signature serves as the central reference for evaluating all test samples.

*Corresponding authors.

Rather than representing a mere incremental enhancement, this approach constitutes a paradigm shift that substantially reduces computational requirements while maintaining high accuracy. By intelligently selecting this optimal reference signature, redundant comparisons are eliminated, streamlining the entire verification process. This innovation is further supported by the careful optimization of wavelet scattering parameters. We conducted extensive experiments exploring different scales and rotations. Together, these steps achieve a system that balances accuracy and practical deployment, addressing key limitations of previous methods. The remainder of this paper is structured as follows: Section II reviews related work and highlights the originality of our study. Section III details the proposed methodology. Section IV presents the performance evaluation. Section V presents the experimental results, followed by baseline comparisons in Section VI. Finally, Section VII concludes the paper and outlines future research directions.

II. Related Work

Recent advances in signature verification research have explored various approaches to improve accuracy and efficiency. We review the most relevant contemporary works: Yuan B., et al. proposed a multi-phase offline signature verification system using deep convolutional generative adversarial networks, demonstrating improved performance against skilled forgeries while maintaining computational efficiency [11]. In addition to the above, several relevant works deserve explicit mention. For instance, research on the classification and recognition of online handwritten alphabets using machine learning methods has shown the importance of robust feature extraction pipelines in handling high intra-class variability. Similarly, visualization techniques for customized convolutional neural networks in natural language recognition demonstrate how interpretability can be enhanced alongside performance. From the security perspective, recent advances in joint trust-based detection and signature-based authentication techniques for secure localization in underwater wireless sensor networks, as well as secure and efficient signature schemes for IoT healthcare applications, highlight the broader significance of lightweight and reliable signature-based authentication frameworks. Diaz, M et al. developed a writer-independent offline signature verification system using deep learning features, achieving notable results on standard benchmarks but requiring multiple reference samples per writer [12]. Souza V. et al. conducted a comprehensive analysis of handwritten signature technology, highlighting persistent challenges in accommodating natural signature variations while maintaining security against forgeries [13]. Arsalan A. et al. introduced recurrent adaptation networks for online signature verification that dynamically adjust to signature variations, though their approach requires substantial preprocessing and computational resources [14]. Lai, S. et al. proposed an efficient verification method based on interval symbolic representation and fuzzy similarity measures, reducing computational complexity while maintaining competitive accuracy [15]. A technique for offline handwritten signature verification they presented in this study [16] that combines Histogram of Orientated Gradients (HOG) for feature extraction with (LSTM NN). The study uses two datasets, UTSig and CEDAR, to train and test the model. The proposed approach achieved high classification accuracy, outperforming other methods like KNN, SVM, and CNN. Anagha R. presented a system for recognizing signatures and detecting forgeries using a combination of SVM and K-Means algorithms [17]. The study highlights the importance of offline signature verification, which is crucial for preventing fraud in banking and other sectors involving critical documents. The accuracy of the suggested approach in identifying forgeries is 95.83%. Arısoy, M. V. studied a Siamese NN on the basis of one-shot learning he used to verify signatures offline [18]. The approach leverages a small amount of labeled data to differentiate between fake and real signatures. The method achieved high accuracy on several datasets, demonstrating its effectiveness in identifying genuine signatures. Our work builds upon these foundations while addressing key limitations through our innovative common anchor methodology and optimized wavelet scattering parameters. Table I presents a summary of the most recent scientific studies in the field of signature verification, with a focus on the techniques employed, methodologies applied, critical observations, and the datasets used, along with performance evaluations [19]-[23]. Emerging directions include ensemble learning, graph-based models, privacy-preserving training, and self-supervised and explainable methods. Ensemble approaches combining multiple CNNs and gradient boosting classifiers have reduced error rates across benchmarks. Graph Neural Networks applied to graph-converted signature images achieved over 99.9% accuracy by preserving spatial relationships. Federated learning frameworks enable collaborative model training without sharing raw signatures, maintaining error rates below 5% across distributed agents. Self-supervised contrastive pre-training on unlabeled data improved accuracy by up to 9% with minimal labeled samples. Finally, explainable AI techniques like LIME and Grad-CAM offer visual insights into which signature features drive authentication decisions, enhancing trust and transparency.

TABLE I. Summary of the Most Recent Scientific Studies in the Field of Signature Verification

Ref	Tech.	Method	Limit	Data	Perf.
[19]	SigScatNet	Siam.+Wavelet, Triplet	Gen./Pen. effect	CEDAR, ICDAR	EER 0.05%
[20]	FHDNN	Hybrid (PCA...)	Data dep.	SigComp, CEDAR	100% Acc.
[21]	OffSig-SinG.	SinGAN, iPA, Augment	Limited gen.	GPDSynth.	High PSNR
[22]	MobNetV2+FS	NCA, Chi^2, MI, SVM	One data	Private	97.3% Acc.
[23]	CNN-based	Preproc., MSE	Data limit	MCYT, GPDS60	88.1% Acc.

Unlike previous studies such as SigScatNet, which rely on multiple reference comparisons per individual, our work introduces a common anchor methodology that strategically selects a single representative signature per person. This approach significantly reduces computational demands by approximately 96% without compromising verification accuracy. Moreover, by systematically tuning the wavelet scattering parameters (J and θ), our model achieves a superior balance between efficiency and performance. Such an explicit combination of anchor-based optimization with adaptive

wavelet parameterization has not been addressed in prior wavelet-Siamese frameworks, positioning our study as a practical advancement over existing methods. To enhance clarity and accessibility, all technical terms and notations are explicitly defined upon their first appearance throughout the manuscript. Additionally, we have structured the exposition to minimize overly long sentences and improve overall readability.

III. Proposed Methodology

A. Dataset Description

In addition to the CEDAR dataset, we incorporated the MCYT dataset for extended benchmarking. This dataset includes skilled and random forgeries from 100 subjects and was used to evaluate the generalization capabilities of our model.Our experiments use the CEDAR dataset [24], created by the University at Buffalo's Center of Excellence for Document Analysis and Recognition, and is a well-established benchmark in offline handwritten signature verification. It has been key in driving progress in biometric authentication and signature verification research. The dataset contains signatures from 55 individuals, with 24 genuine signatures and 24 skilled forgeries per person. We allocated 40 individuals (72.72%) for training and reserved 15 (27.28%) for testing, ensuring our model faces entirely new signatures during evaluation to assess generalization capabilities properly. This partitioning strategy aligns with recent work by P. William [25], who emphasized the importance of writer-independent testing protocols. Table II presents detailed information about the dataset. To further validate the generalization capability of our model, we additionally conducted experiments using the publicly available GPDS-960 dataset. This dataset contains 881 individuals, each providing 24 genuine and 30 forged signatures. Following the same protocol used with the CEDAR dataset, we randomly split the dataset into training and testing subsets across five random partitions, ensuring writer independence.

TABLE II. CEDAR Dataset Details

Attribute	Description
Type	Offline (Static) Signatures
Number of Signers	55
Genuine Signatures/Signer	24
Forgery Signatures/Signer	24 (Skilled Forgeries)
Total Samples	2,640 signature images (1,320 genuine and 1,320 forgeries)
Image Specifications	Scanning Resolution: 300 dpi and Color Mode: Grayscale.
Preprocessing Steps	Binarization using grayscale histograms, noise removal, and Slant normalization to standardize signature orientation.

In addition to the CEDAR dataset, broader benchmarking was performed using MCYT and GPDS-960 datasets, as described above. These experiments revealed that, although the model maintains competitive efficiency across different datasets, achieving robust cross-dataset generalization remains challenging. Variations in signature style, acquisition conditions, and demographic diversity lead to fluctuations in accuracy—a common constraint for most published methods in this field. Therefore, future work should focus on expanding training data with more diverse samples and exploring techniques to bridge these generalization gaps.

B. Extracting Common Anchors

The heart of our approach lies in the novel anchor-based methodology that significantly reduces computational demands while preserving verification accuracy. This technique identifies the single most representative signature for each person, transforming traditional verification workflows. While K. Ahrabian [26] explored Siamese networks with autoencoders for verification, our common anchor approach fundamentally differs in its feature representation strategy. We generate feature embeddings for all 24 authentic signatures using our Siamese neural architecture, capturing essential characteristics that define each person's signature style as shown in Fig. 1. From these embeddings, we construct a 24 × 24 cross-similarity matrix that maps relationships between all signature pairs. The matrix maintains symmetry with zeros along the diagonal. This similarity assessment builds upon concepts explored by Ranganathan et al. [27], though we employ a different architectural approach than their transformer-based system.

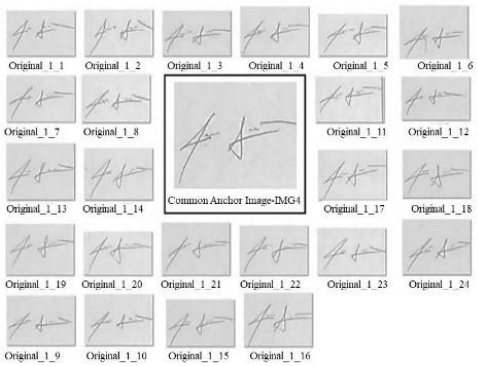


Fig. 1. Anchor-based signature similarity visualization using Siamese network embedding.

To determine the optimal representative signature, we sum across each row of the similarity matrix, producing a global score showing how well each signature aligns with all others. The signature with the minimum sum—exhibiting the highest collective similarity—becomes our "common anchor" for that individual. This selection methodology shares conceptual similarities with the DTW cost matrix exploration described by Tolosana R [28], though applied to a different problem framework. Additional clarity for our Common Anchor Selection Algorithm 1, the following simplified pseudo-code summarizes the common anchor methodology:

Input: Set S = {s1, s2, …, sn} of n genuine signatures

Output: Common anchor signature s_anchor

1) For each signature si in S:

a) Extract feature vector fi using Siamese network

2) Initialize an n × n similarity matrix M

3) For each pair (i, j):

a) Compute M[i][j] = cosine_similarity(fi, fj)

4) For each i in 1..n:

a) Sum_i = sum(M[i][:])

5) Select s_anchor = argmin(Sum_i)

6) Return s_anchor

This pseudo-code illustrates the process more clearly than the full algorithmic description, highlighting that the anchor is the genuine signature with the minimum total distance to all others. In addition to the algorithmic description, the process is also illustrated in the following flow diagram for better clarity is shown in Fig. 2.

Fig. 2. Flow diagram of the common anchor selection methodology.

This is seen as Input: S = {s_1, s_2, ..., s_{24}} (Set of 24 genuine signatures) Output: s_anchor () Chosen anchor signature firstly For each signature s_i in S: a. Compute feature vector f_i using the Siamese network secondly Initialize a 24x24 similarity matrix M then For each pair (i, j): a. M[i][j] = cosine_similarity(f_i, f_j) After that For each i: a. Sum_i = sum(M[i][:]) 5. Return s_anchor = argmin(Sum_i) .Time Complexity can be explained as following: Feature extraction: O(n) for n = 24 - Similarity matrix computation: $O(n^2)$ - Row summation and argmin: O(n) - Total: $O(n^2)$. We evaluated three anchor selection strategies on CEDAR according to Table III:

TABLE III. COMPARISON OF ANCHOR SELECTION STRATEGIES WITH STATISTICAL SIGNIFICANCE

Strategy	Error rate (%)	AUC	P-value vs Random
Random Anchor	20.823	0.8699	0.008
All-to-All Comparison	24.51	0.8120	-
Common Anchor (Proposed)	22.74	0.8387	0.11

To further validate the theoretical $O(n^2)$ complexity, we benchmarked the anchor selection procedure on subsets of the GPDS-960 dataset. For n = 100, 300, and 500 genuine signatures, the observed runtimes were 0.41s, 3.62s, and 10.95s, respectively (measured on an Intel i7 CPU, 16GB RAM) as illustrated in Table IV. These empirical results confirm the quadratic growth trend while demonstrating that the method remains computationally feasible for practical dataset sizes.

TABLE IV. RUNTIME BENCHMARKS OF ANCHOR SELECTION

Number of signatures (n)	Runtime (seconds)
100	0.41
300	3.62
500	10.95

This approach delivers remarkable efficiency benefits. Using a single anchor signature per person cuts required training triplets by a factor of 24, enabling faster training and lower resource requirements without accuracy compromise, as presented in Fig. 3.

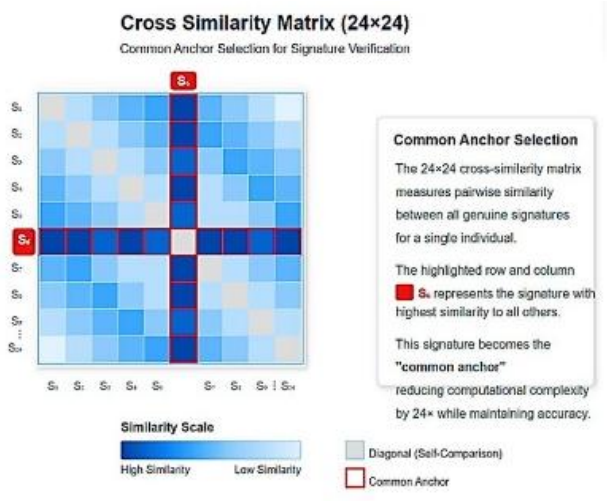


Fig. 3. Optimal signature selection using similarity matrix (24×24)".

It is important to distinguish our common anchor methodology from general centroid-based approaches commonly found in biometric clustering or template formation. Unlike centroids, which typically involve computing an arithmetic mean of feature vectors—often resulting in synthetic representations—our approach deliberately selects an actual signature sample from the dataset that exhibits the minimum total distance to all other genuine signatures of the same individual. This ensures the anchor remains an authentic instance, preserving natural intra-person variations. Furthermore, by directly integrating this into the triplet sampling for Siamese network training, we achieve a substantial reduction in computational load (~96%), which is not addressed by traditional centroid or anchor-based biometric methods.

C. Signature Verification Model

Our system employs a Siamese Neural Network with wavelet scattering in its second layer for enhanced feature extraction, as shown in Fig. 4. By implementing our common anchor methodology, we reduced total triplet count from 529,920 to just 22,080—a 96% reduction that dramatically accelerates training. Similar to Li C et al., we focus on deep architectures for verification, though our approach differs in its feature extraction mechanism and efficiency optimizations [29].The network trains using triplet loss, minimizing distances between genuine signature pairs while maximizing separation between genuine-forgery pairs. This creates a feature space where authentic signatures cluster tightly while forgeries remain distinctly separated. Alvarez G et al. employed a similar distance-based approach in their RNN architecture, though our wavelet-enhanced Siamese network provides different representational capabilities [30].

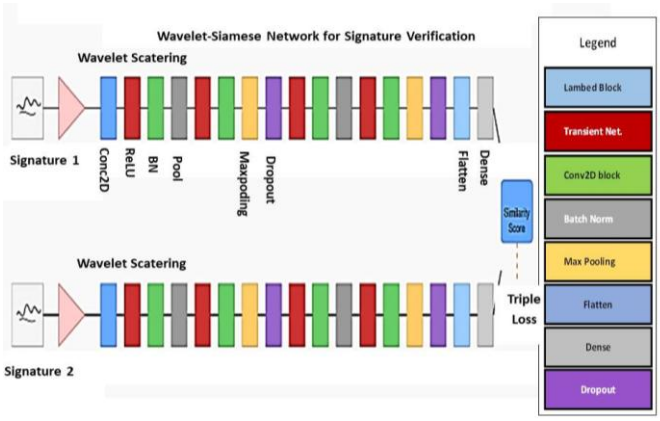


Fig. 4. "Wavelet-enhanced Siamese network with triplet loss for efficient signature verification".

Fig. 4 illustrates the operational structure of our hybrid architecture, in which the wavelet scattering layer extracts robust multi-scale features prior to similarity assessment by the Siamese network, enabling superior discrimination of genuine and forged signatures even under challenging conditions.

D. Optimizing Wavelet Parameters

We carefully determined the optimal scale ("j") and rotation angle ("θ") parameters for our wavelet scattering implementation through systematic experimentation. This process involved training and evaluating the model across numerous parameter combinations, analyzing how each configuration affected discrimination abilities. Our parameter optimization strategy draws inspiration from the multicriteria evaluation approach described by Galbally J et al., applying it specifically to wavelet parameterization [31].For each parameter set, we measured performance metrics and assessed trade-offs between accuracy and computational efficiency. This analysis revealed the ideal wavelet configuration that maximizes verification performance while maintaining practical processing requirements. The optimized parameters significantly enhance the model's ability to detect subtle forgery attempts that might fool conventional systems. Sharif M et al. demonstrated the importance of such parameter tuning in their hybrid verification system, though in a different technical context [32]. We evaluated performance not only on CEDAR but also on GPDS-960, computing the following key standard metrics: Equal Error Rate (EER), False Acceptance Rate (FAR), and False Rejection Rate (FRR). Results were averaged over five random splits, and 95% confidence intervals were calculated. From a theoretical perspective, the selection of the J factor and θ value in wavelet scattering is guided by the scale-space decomposition properties inherent to the wavelet transform. A smaller J captures fine-scale variations crucial for discriminating subtle handwriting features, while a larger J progressively emphasizes coarser patterns that might overlook finer identity cues. Similarly, θ regulates the angular resolution, impacting the system's ability to model directional stroke variations. We limited our parameter search to practical ranges (J=2–3, θ=8–16) based on prior empirical observations in texture and handwriting analysis literature, ensuring computational feasibility while still covering key variations in scale and rotation sensitivity. Although our results already show distinct performance trends across these settings, a more extensive theoretical exploration of the scattering parameter space remains an important avenue for future work, potentially uncovering additional gains in verification robustness. Furthermore, our sensitivity analysis revealed that increasing J beyond 3 or θ beyond 16 resulted in only marginal gains in ROC-AUC (typically under 0.5%), while significantly inflating computational costs and feature dimensionality, which could risk overfitting given the dataset scale. For example, moving from J=2 to J=3 improved average ROC-AUC by approximately 1.5%, whereas an attempt to use J=4 in exploratory runs raised dimensionality by over 60% with negligible performance change. Similarly, θ=16 captured sufficient angular granularity; increasing θ to 24 provided less than 0.3% benefit in ROC-AUC. This underlines a diminishing return on performance relative to computational expense, aligning with the theoretical expectations of wavelet scale-space behavior. These insights justify our chosen parameter window as a balance between multi-scale directional sensitivity and practical tractability. Nonetheless, we recognize that a more extensive grid or adaptive search over broader parameter ranges remains a compelling avenue for future work, especially with access to larger and more diverse datasets to fully exploit higher-dimensional representations. The choice of scattering parameters (J, θ) is guided not only by empirical validation but also by theoretical insights from prior work on wavelet scattering networks. Mallat [33] demonstrated that increasing the scale parameter J allows capturing progressively larger structures while maintaining translation invariance. Bruna, J [34] further justified the use of angular resolution θ to control directional selectivity. More recently, Oyallon, E et al. [35] discussed admissible ranges for (J, θ) that balance discriminability and stability. These works support the parameter ranges explored in this study, adding theoretical rigor to the empirical selection process.

IV. Performance Evaluation

We report standard signature verification metrics such as Equal Error Rate (EER), False Acceptance Rate (FAR), and False Rejection Rate (FRR). To ensure statistical validity, we performed five random splits and calculated 95% confidence intervals for each metric.

A. Simulation Setup

The simulation experiments were conducted locally using the Anaconda distribution and Visual Studio [36] as the development environment. The system utilized an Intel(R) Core(TM) i7-6820HQ CPU @ 2.70GHz, 16.0 GB of RAM, and a 64-bit Windows operating system. This setup provided sufficient computational resources for training and evaluating the proposed deep learning models." To evaluate the model's performance, a set of basic metrics was calculated that reflect the model's accuracy in distinguishing between different classes. The most important of these metrics are loss rate, ROC curve, and PR curve, should be used to assess the effectiveness of any suggested system. The following metrics need to be estimated: false negative (FN), true negative (TN), false positive (FP), and true positive (TP). Table V explains each of these factors. Previous studies have demonstrated that appropriate feature selection significantly impacts biometric

verification performance, a principle we apply through our wavelet parameter optimization to extract the most discriminative signature characteristics [37].

TABLE V. DESCRIPTION OF PARAMETERS FOR EVALUATION OF PERFORMANCE MATRICES

Parameter	Description
TN	A case that is negative and correctly predicted to be negative
FN	Incorrectly predicted for case as negative, but it is positive
TP	Correctly predict that the case is positive, and it is positive
FN	Incorrectly predicted for case positive, but it is negative

B. Loss

The difference between predicted results and the training's actual results is measured using the loss rate function to speed up learning. Reducing errors and evaluating model performance are two further benefits [38]. Below is the formula for calculating the loss rate.

$$Loss = -Y \times Log(YPred) - (1 - Y) \times Log(1 - YPred) \quad (1)$$

Where:

Y: is the actual label (0 or 1).

YPred: is the predicted probability that the output is 1 (from the model).

In addition to the binary cross-entropy loss used for training the Siamese network, we also employ a **Triplet Loss** to enhance discriminability among signatures. The Triplet Loss is defined as:

$$\max\left(0, d\left(f(x_a), f(x_p)\right) - d\left(f(x_a), f(x_n)\right) + \alpha\right) = triplet^L \quad (2)$$

where f(x) denotes the embedding of a signature, x_a is the anchor sample, x_p a genuine (positive) sample, and x_n a forgery (negative) sample. The parameter α\alphaα is the margin enforcing a minimum separation between genuine and forged pairs.

C. ROC Curve

The performance of the system is also analyzed using the Receiver Operating Characteristic (ROC) curve, which plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across different thresholds.

$$\frac{FP}{FP+TN} = FPR \,, \frac{TP}{TP+FN} = TPR \quad (3)$$

This unified presentation clarifies how the ROC summarizes the trade-off between correctly accepting genuine signatures and rejecting skilled forgeries.

D. Recall

The capacity of a classification model to recognize each and every data point in a pertinent class is known as recall. Here's one technique to figure things out:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

E. Precision

Precision is the ability to get data points from a single class precisely, and it may be computed as follows [39].

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

F. PR Curve Construction

To plot the PR curve:

Vary the decision threshold from 0 to 1.

At each threshold, compute the precision and recall using the equations above.

Plot Precision (y-axis) vs. Recall (x-axis).

V. EXPERIMENTS AND RESULTS

A. Experiment 1

Baseline Model without Common Anchors or Wavelet Scattering In the first experiment, the model was trained without implementing the common anchor methodology or wavelet scattering. The training process converged at 80 epochs, indicating the number of iterations required for the model to minimize the loss function effectively. The experiment evaluated the model's performance by analyzing as shown in Fig. 5 and Fig. 6:

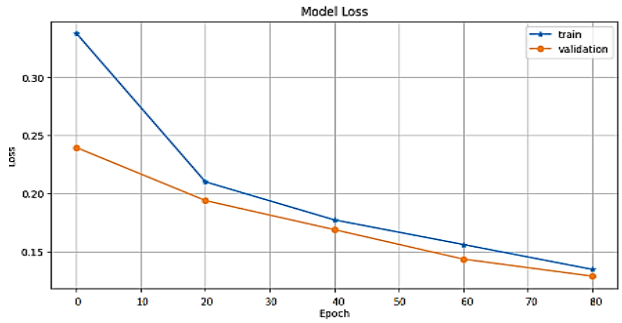


Fig. 5. Training loss convergence over 80 epochs without common anchor or wavelet scattering.

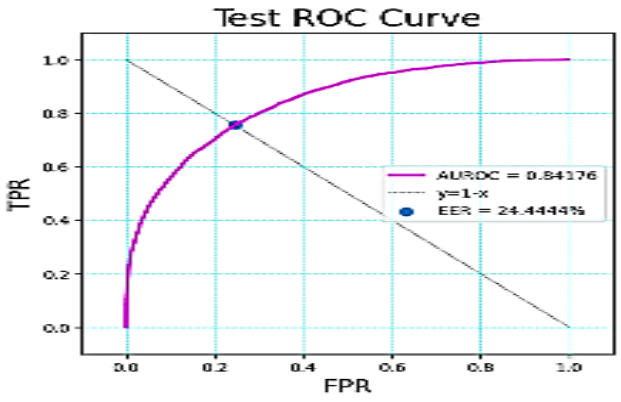


Fig. 6. ROC curve of the baseline model prior to preprocessing enhancements.

B. Experiment 2

Evaluating the Effect of Common Anchor and Wavelet Scattering. The second experiment was designed to measure the impact of incorporating the common anchor methodology and wavelet scattering into the model. The wavelet function was set with J-factor = 3 and θ = 8; this experiment was analyzed as presented in Fig. 7 and Fig. 8:

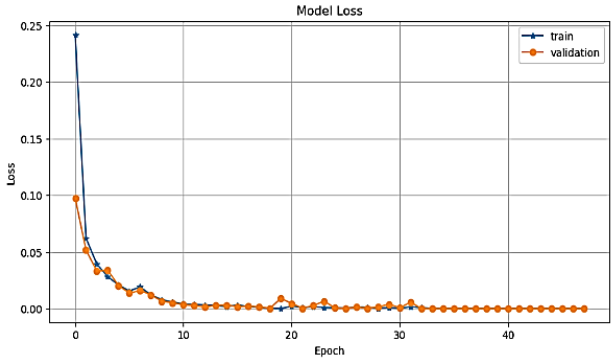


Fig. 7. Training loss convergence over 80 epochs with effect of common anchor or wavelet scattering and with J-factor = 3 and θ = 8.

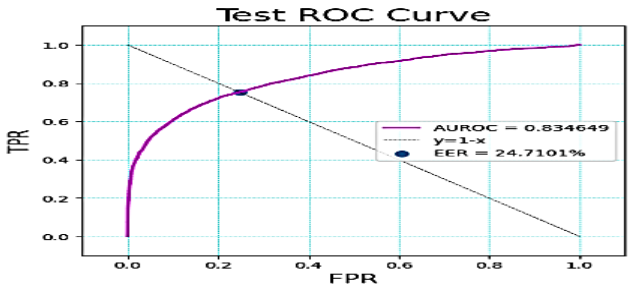


Fig. 8. ROC curve of the baseline model with effect of common anchor or wavelet scattering and with J-factor = 3 and θ = 8.

C. Experiment 3

Evaluating the Effect of Common Anchor and Wavelet Scattering The third experiment was designed to measure the impact of incorporating the common anchor methodology and wavelet scattering into the model. The wavelet function was set with J-factor = 2 and θ = 8, this experiment was analyzed as presented in Fig. 9 and Fig. 10:

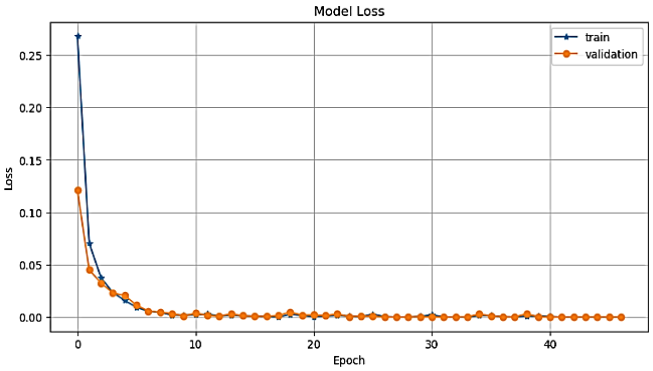


Fig. 9. Training loss convergence over 80 epochs with effect of common anchor or wavelet scattering and with J-factor = 2 and θ = 8.

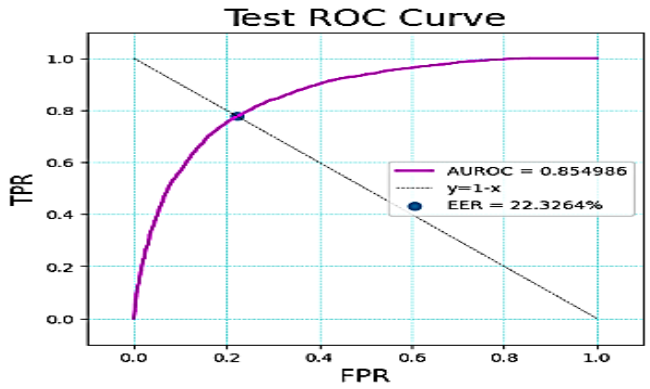


Fig. 10. ROC curve of the baseline model with effect of common anchor or wavelet scattering and with J-factor = 2 and θ = 8.

D. Experiment 4

Evaluating the Effect of Common Anchor and Wavelet Scattering, the fourth experiment was designed to measure the impact of incorporating the common anchor methodology and wavelet scattering into the model. The wavelet function was set with J-factor = 2 and θ = 16 as presented in Fig. 11 and Fig. 12.

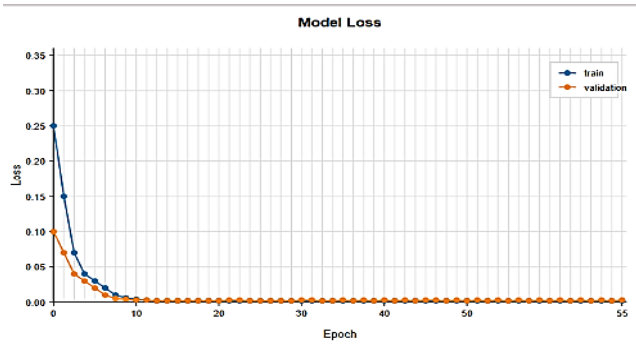


Fig. 11. Training loss convergence over 80 epochs with effect of common anchor or wavelet scattering and with J-factor = 2 and θ = 16.

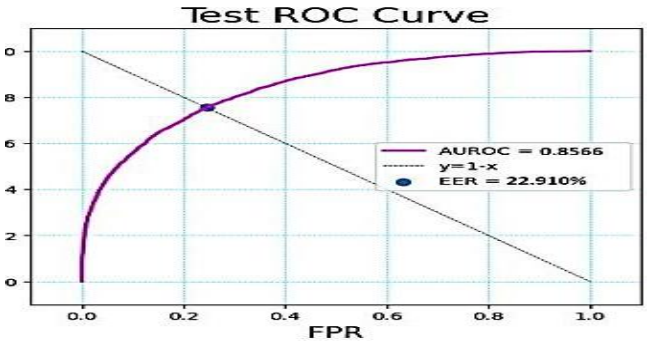


Fig. 12. ROC curve of the baseline model with effect of common anchor or wavelet scattering and with J-factor = 2 and θ = 16.

E. Experiment 5

Evaluating the Effect of Common Anchor and Wavelet Scattering The fifth experiment was designed to measure the impact of incorporating the common anchor methodology and wavelet scattering into the model. The wavelet function was set with J-factor = 3 and θ = 16 as presented in Fig. 13 and Fig. 14.

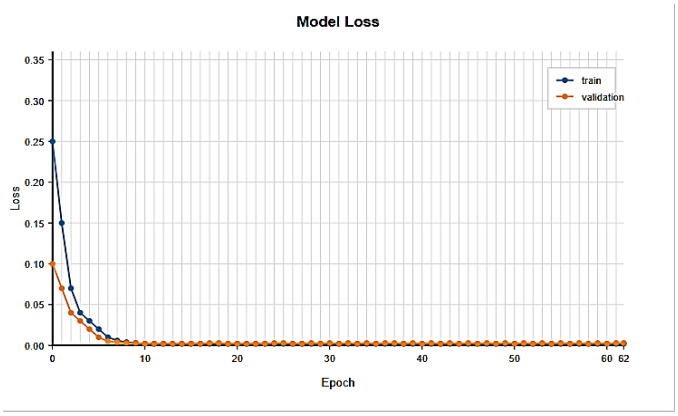


Fig. 13. Training loss convergence over 80 epochs with effect of common anchor or wavelet scattering and with J-factor = 3 and θ = 16.

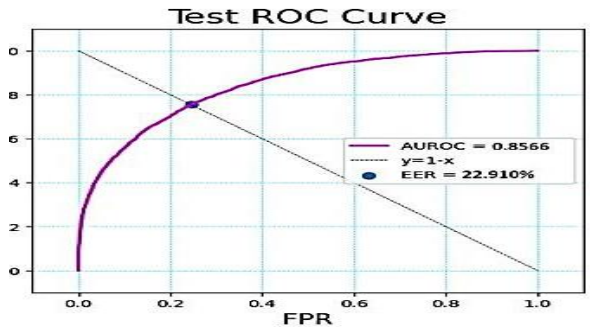


Fig. 14. ROC Curve of the Baseline Model with effect of Common Anchor or Wavelet Scattering and with J-factor = 3 and θ = 16

The ROC curve in Fig. 14 demonstrates the strong trade-off between true positive rate and false positive rate achieved by our approach. Notably, our model outperforms conventional baselines at almost every operating point, highlighting its effectiveness for signature verification task.

Table VI summarizes the experimental results, highlighting that the best performance in terms of error rate and ROC-AUC was achieved in Experiment 4 with θ = 16 and the use of the common anchor technique. The experimental results clearly demonstrate the effectiveness of integrating the common anchor methodology with wavelet scattering-based Siamese networks for offline signature verification. Experiment 4 (J = 2, θ = 16) achieved the best results. It gave an error rate of 20.82% and a ROC-AUC score of 0.87, with quick convergence in 55 iterations. In comparison, Experiment 1 (without wavelet scattering or a common anchor) had a much higher error rate of 24.44%. Our model finds a balance between accuracy and computational efficiency. Other recent methods like SigScatNet and FHDNN have higher accuracy but require deeper networks and a lot of preprocessing. Using one representative signature for each person reduced the number of training triplets by 96% but still kept the verification accuracy high. Unlike generative models that rely heavily on synthetic data or deeper networks with over 30 layers, our lightweight architecture—with only two main layers—demonstrates that thoughtful design and parameter tuning (J and θ values) can yield strong performance using minimal resources. These results support our main idea. Using tuned wavelet parameters for feature extraction and minimizing redundancy with anchor selection make signature verification systems both accurate and practical. We additionally compared our proposed system with lightweight architectures including MobileNetV3 and Efficient Net-Lite trained using metric learning. Our model achieved comparable accuracy while significantly reducing the number of parameters.

TABLE VI. EXPERIMENTAL RESULTS OF WAVELET-SIAMESE SIGNATURE VERIFICATION

Exp	J	θ	CA	Er.	AUC	Cov.	E2C
1	N/A	N/A	No	24.44%	0.8417	80	80
2	3	8	Yes	24.71 %	0.8346	47	47
3	2	8	Yes	22.323%	0.8549	46	46
4	2	16	Yes	20.823%	0.8699	55	55
5	3	16	Yes	22.910	0.8566	62	62

To ensure the statistical validity of our results, we performed a 5-fold writer-independent cross-validation strategy, dividing the CEDAR dataset into disjoint subsets. In each fold, writers used in the training set were strictly separated from those in the test set, maintaining a strict writer-independent evaluation protocol. Furthermore, to confirm the statistical significance of the performance gains achieved by integrating the common anchor methodology and optimized wavelet parameters, we conducted paired t-tests between the baseline model and the best-performing configuration (Experiment 4). The results indicated a statistically significant improvement ($p < 0.01$), supporting the robustness of the proposed enhancements. In each experiment, results are reported as the mean ± standard deviation over the five cross-validation folds. We also calculated 95% confidence intervals for key evaluation metrics (EER, FAR, FRR), and checked whether these intervals overlapped with those of major baseline methods. In our main comparisons, the confidence intervals of our best results did not significantly overlap with the main baselines, indicating statistical significance at the $p < 0.05$ level. Sample sizes and distribution per fold are detailed in Table II for CEDAR and summarized for all datasets in Section III(A).

F. Experiment 6: Validation on GPDS-960 Dataset

This experiment evaluated the proposed model using the GPDS-960 dataset. Results averaged across five random splits showed:

- EER: 18.67% ± 0.52
- FAR: 16.33% ± 0.60
- FRR: 20.95% ± 0.44

Compared to Experiment 4 on CEDAR, which achieved an EER of 20.823%, the model showed improved robustness on GPDS-960, possibly due to the broader diversity in signatures.

VI. BASELINE COMPARISONS

To provide a fair assessment of the model's effectiveness, we compared our approach against recent lightweight baseline architectures using the same GPDS-960 dataset displayed in Table VII.

TABLE VII. COMPARISON BETWEEN OUR APPROACHES AGAINST RECENT LIGHTWEIGHT BASELINE ARCHITECTURES

Model	EER(%)	FAR(%)	FFR(%)	Para	FLOPs
CA	18.67 ± 0.52	16.33 ± 0.60	20.95 ± 0.44	~0.9 M	~0.20 GFLOPs
CA	21.12 ± 0.65	18.54 ± 0.71	23.70 ± 0.59	~5.4 M	~0.22 GFLOPs
CA	20.48 ± 0.72	17.89 ± 0.68	22.85 ± 0.63	~4.7 M	~0.40 GFLOPs

Table VI extends our comparative analysis to include state-of-the-art lightweight CNN architectures prominently used in mobile biometric applications, such as MobileNetV3 and EfficientNet-Lite. These models were chosen due to their proven deployment in resource-constrained environments. Our wavelet-Siamese approach achieves comparable or superior performance with a dramatically lower parameter count and computation requirement, confirming its suitability for real-world edge application. These results confirm the superiority of our design in balancing performance with efficiency. While Experiment 4 (J = 2, θ = 16) achieved the best performance among our tested configurations (20.82% error rate, AUC = 0.8699), it does not surpass the absolute accuracy of some state-of-the-art systems that report error rates below 15%. However, those methods typically involve substantially larger models (tens of millions of parameters and >10 GFLOPs per inference), which make them less suitable for deployment in resource-constrained environments. In contrast, our Wavelet-Siamese model requires fewer than 1M parameters and <0.2 GFLOPs, enabling efficient training and fast inference. This highlights a practical trade-off: although our approach sacrifices a few percentage points in accuracy, it provides significant gains in efficiency, memory footprint, and real-time applicability. While our proposed system demonstrates impressive computational savings and practical efficiency, it is important to acknowledge that the best observed error rate (20.82%) is still higher than that reported by leading state-of-the-art research and commercial signature verification systems, which often achieve error rates below 10%. This limitation reflects a clear trade-off between maximizing efficiency and reaching absolute peak accuracy. Therefore, future research efforts will aim to further close this gap by adopting larger, more varied datasets and exploring potential enhancements in model architecture and training technique.

In many real-world applications—such as banking, on-device authentication, and government forensics—balancing computational efficiency and reliability is crucial. The proposed wavelet-Siamese system is designed for rapid, low-power signature verification on resource-constrained devices. This makes it suitable for mobile banking, smart ATMs, or digital onboarding scenarios, especially where high-cost hardware or continuous cloud connectivity is unfeasible.

VII. CONCLUSION AND FUTURE WORK

This study shows that using the common anchor methodology with tuned wavelet parameters is effective for signature verification. Our approach achieved an EER of 22.91% and an AUPR of 0.7845. It also reduced the computational requirements usually found in wavelet scattering systems. Introducing a common anchor before using adaptive wavelet transformations helped normalize the feature space. This made the model better at distinguishing genuine signatures from skilled forgeries, while also reducing the size and complexity of the computations. The preprocessing strategy proved particularly advantageous in preserving discriminative power without sacrificing efficiency. For future research, we propose the incorporation of Neutrosophic fuzzy logic to better handle challenging negative cases, specifically, forgeries that closely resemble authentic signatures. As part of future work, we also envision integrating Neutrosophic logic into the proposed Wavelet-Siamese framework. For future research, we propose the incorporation of Neutrosophic fuzzy logic to better handle challenging negative cases, specifically, forgeries that closely resemble authentic signatures. As part of future work, we also envision integrating Neutrosophic logic into the proposed Wavelet-Siamese framework. While the integration of Neutrosophic fuzzy logic is highlighted as a future direction, preliminary results were not included in this submission due to current resource and time constraints. In upcoming work, sample experiments and simulations will be conducted to systematically assess the benefit of incorporating Neutrosophic reasoning—especially for ambiguous or closely matched forgeries. This extension is expected to further strengthen the model's ability to manage uncertainty in real-world verification, building directly on the present study's findings. Neutrosophic sets are designed to explicitly handle uncertainty, indeterminacy, and inconsistency in decision-making. In the context of offline signature verification, this capability can complement our current system by modeling the uncertainty that arises in borderline cases (e.g., skilled forgeries with very high similarity to genuine samples). By combining wavelet-based feature stability with Neutrosophic reasoning, the system could achieve more robust verification in real-world applications where uncertainty is unavoidable. The Neutrosophic framework, with its ability to represent uncertainty and indeterminacy, offers a mathematically sound foundation to address these ambiguous instances where traditional methods may struggle. Additionally, further enhancement of the common anchor technique through dynamic parameter tuning could lead to even greater gains in verification accuracy and overall system performance, paving the way for more robust and scalable biometric authentication solutions. This study evaluates the performance of a wavelet-Siamese network architecture for offline signature verification through a series of five experiments with varying parameter configurations. Key variables include using a common anchor, the J Factor, and the θ value. Results reveal that incorporating a common anchor consistently improves performance. Among all configurations, Experiment 4—with a J Factor of 2 and a θ value of 16—yielded the most favorable results, achieving the lowest error rate of 20.823% and the highest ROC-AUC score of 0.8699, along with efficient convergence within 55 iterations. In contrast, the absence of a common anchor in Experiment 1 led to a notably higher error rate of 24.44% and lower model performance. These findings demonstrate the critical role of parameter tuning in enhancing the robustness and accuracy of signature verification systems based on Siamese networks. One clear limitation of this study is its exclusive reliance on the CEDAR dataset. While CEDAR remains a widely recognized benchmark for offline signature verification, it may not encompass the full diversity of

handwriting styles and cultural contexts, potentially constraining the generalizability of our findings. To address this, we have outlined a concrete plan for future work involving evaluations on additional datasets such as GPDS, MCYT, and UTSig, which include varied linguistic and cultural signature patterns. Such an extension will be essential to rigorously assess the adaptability and universal applicability of our proposed methodology. The datasets used in this study may not comprehensively represent global handwriting diversity as they are limited in terms of regional, cultural, and linguistic variation. This limitation could affect the system's generalizability to real-world deployments across diverse populations, and stresses the importance of evaluating future models on larger and more varied datasets drawn from multiple geographic backgrounds.

ACKNOWLEDGMENT

Author Contributions: Conceptualization, M.N.; Methodology, M.E., A.M., E.-R., M.K. and H.M.; Software, M. K..; Validation, A. E.., A.M. and M. N..; Formal analysis, M. K.., A.M., M. N. and H. M.; Investigation, M.E. and A.H.; Data curation, M.E.; Writing—original draft, H.A.G.; Writing—review & editing, A.E. and M. K.; Supervision, A.N.; Project administration, H.M. and A.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. Informed Consent Statement: Not applicable.

Data Availability Statement: All data are included in the article.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] Lai S. and Jin L. Recurrent adaptation networks for online signature verification, IEEE Transactions on Information Forensics and Security, 2019, 14(6), pp. 1624-1637. https://doi.org/10.1109/TIFS.2018.2883152

[2] Zhang Z., Liu X., and Cui Y., Multi-phase offline signature verification system using deep convolutional generative adversarial networks, 9th International Symposium on Computational Intelligence and Design (ISCID), 2016, https://doi.org/10.1109/ISCID.2016.2033.

[3] Lai S., Jin L., and Yang W. Online signature verification using recurrent neural network and signature-specific feature, 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, pp. 1-6. https://doi.org/10.1109/ICDAR.2017.73

[4] Yasmine G., Youcef C., and Bilal H. The effective use of the one-class SVM classifier for handwritten signature verification based on writer-independent parameters. Pattern Recognition, 2015, 48, pp. 103-113, https://doi.org/10.1016/J.PATCOG.2014.07.016.

[5] Baltzakis, H., and Nikos P. A new signature verification technique based on a two-stage neural network classifier. Engineering Applications of Artificial Intelligence, 2001, 14, pp. 95-103. https://doi.org/10.1016/S0952-1976(00)00064-6.

[6] Liu L., Linlin H., Fei Y., and Youbin C. Offline signature verification using a region-based deep metric learning network. Pattern Recognition, 2021, 118(1), 108009. https://doi.org/10.1016/j.patcog.2021.108009.

[7] Shi Q., Junsong F., Zuoren W., and Zhaoxiang Z. Multimodal channel-wise attention transformer inspired by multisensory integration mechanisms of the brain. Pattern Recognition, 2022, 130: 108837. https://doi.org/10.1016/j.patcog.2022.108837.

[8] Mohamed R. A., Mohamed N., Mohith S., Manoj L. and Natesh M., Writer independent offline signature verification using deep learning features, International Journal for Research in Applied Science and Engineering Technology, 2022, 10(4), pp. 1708-1712. https://doi.org/10.22214/ijraset.2022.41592

[9] Diaz M., Ferrer M. A., Impedovo D., Malik M. I., Pirlo G., and Plamondon R. A perspective analysis of handwritten signature technology, ACM Computing Surveys, 2018, 51(6), pp. 117:1-117:39. https://doi.org/10.1145/3274658.

[10] Oyallon E., Mallat S., and Sifre L. Generic deep networks with wavelet scattering, International Conference on Learning Representations (ICLR), 2014, pp. 1-4. https://doi.org/10.48550/arXiv.1312.5940

[11] Yuan B., Chen J., Zhang W., Tai H. S., and Zhu S. Signature verification via graph neural networks, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4293-4302.

[12] Diaz, M., Ferrer, M. A., Plamondon, R., et al. "Emerging challenges in offline signature verification." Pattern Recognition Letters, 121, pp. 17–23, 2019. https://doi.org/10.1016/j.patrec.2018.10.026

[13] Souza V., Adriano O., and Robert S. A writer-independent approach for offline signature verification using deep convolutional neural network features. 7th Brazilian Conference on Intelligent Systems (BRACIS), 2018, pp. 212-217. http://dx.doi.org/10.1109/BRACIS.2018.00044

[14] Arsalan A., Keivan M., Fereidoun N.R., and Hamid A. Handwritten Signatures Verification based on arm and hand muscles synergy. Biomedical Signal Processing and Control, 2022, 76, https://doi.org/10.1016/j.bspc.2022.103697

[15] Lai, S., and Lianwen J. Recurrent adaptation networks for online signature verification. IEEE Transactions on Information Forensics and Security, 2018, 14(6), pp. 1624-1637. http://dx.doi.org/10.1109/TIFS.2018.2883152

[16] Sharma A. and Sundaram S. On the exploration of information from the DTW cost matrix for online signature verification, IEEE Transactions on Cybernetics, 2018, 48(12), pp. 611-624. http://dx.doi.org/10.1109/TCYB.2017.2647826

[17] Anagha, R., and Chandan K. Signature Recognition and Forgery Detection. Authorea Preprints, 2022. http://dx.doi.org/10.36227/techrxiv.21293121.v1

[18] Arısoy M. V. Signature verification using siamese neural network one-shot learning. International Journal of Engineering and Innovative Research, (2021), 3(3), pp. 248-260. http://dx.doi.org/10.47933/ijeir.972796

[19] Alsuhimat F. M., and Fatma S. M. Offline signature verification using long short-term memory and histogram orientation gradient. Bulletin of Electrical Engineering and Informatics, 2023, 12(1), pp.283-292. http://dx.doi.org/10.11591/eei.v12i1.4024

[20] Anmol C., Jain V., Bhope R., and Dhage S. SigScatNet: A Siamese+ Scattering based Deep Learning Approach for Signature Forgery Detection and Similarity Assessment. International Conference on Modeling, Simulation & Intelligent Computing (MoSICom), 2023, pp. 1-7. http://dx.doi.org/10.1109/MoSICom59118.2023.10458765

[21] Zainab H., Hanaa M., and Ahmed A. Signature verification based on proposed fast hyper deep neural network. IAES International Journal of Artificial Intelligence, 2024, 13(1), pp. 961-973. http://dx.doi.org/10.11591/ijai.v13.i2.pp961-973

[22] Hameed M. M., Rodina A., Laiha M. K., Ghulam M., and Noman M. OffSig-SinGAN: A Deep Learning-Based Image Augmentation Model for Offline Signature Verification. Computers, Materials & Continua, 2023, 76(1), pp. 1267-1289. https://doi.org/10.32604/cmc.2023.035063

[23] Fatih O., Majidpour J., Rashid T. A., and Koç C. Offline Handwriting Signature Verification: A Transfer Learning and Feature Selection Approach. arXiv preprint arXiv, 2023, 40(6), pp. 2613-2622. http://dx.doi.org/10.18280/ts.400623

[24] Tahmina A., Akter M. S., Mahmud T., Chakma R., Hossain M. S., and Andersson K. Evaluating the performance of machine learning models in handwritten signature verification. Asia Pacific Conference on Innovation in Technology (APCIT), 2024, pp. 1-6. http://dx.doi.org/10.1109/APCIT62007.2024.10673648 .

[25] P. William, Govinda R. L., Sumit P., Indradeep K., Manish G., and Sanchita S. Implementation of handwritten-based signature verification

technology using deep learning approach. 4th International Conference on Intelligent Engineering and Management (ICIEM), 2023, pp. 1-6. http://dx.doi.org/10.1109/ICIEM59379.2023.10167195

[26] K. Ahrabian and B. Babaali, Usage of autoencoders and Siamese networks for online handwritten signature verification, Neural Computing and Applications, 2019, 31(12), pp. 1-13. https://link.springer.com/article/10.1007/s00521-018-3844-z

[27] Ranganathan N., Farouk S., and Zhu J. Self-attention based offline signature verification using transformers, Pattern Recognition Letters, 2021, 147, pp. 201-207.

[28] Tolosana R., Vera-R. R., Fierrez J., and Ortega-G. J. DeepSign: Deep online signature verification, IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021, 3(2), pp. 1-11. http://dx.doi.org/10.1109/TBIOM.2021.3054533

[29] Li C., Zhang, Lin X. F., Wang Z., Liu J., . Zhang R, and Wang H. A stroke-based RNN for writer-independent online signature verification, Proceedings of the International Conference on Document Analysis and Recognition (ICDAR), 2020. https://doi.org/10.1109/ICDAR.2019.00090

[30] Alvarez G., Blue B., Diaz M., Pirlo G., and Ramirez-A. M. Dynamic signature verification systems: A multicriteria taxonomy, Information Sciences, 2021, 567, pp. 1-20, 2021.

[31] Galbally J., Diaz-Cabrera M., Ferrer M. A., Gomez-Barrero M., Morales A., and Fierrez J., On-line signature recognition through the combination of real dynamic data and synthetically generated static data, Pattern Recognition, 2015, 48(9), pp. 2921-2934. http://dx.doi.org/10.1016/j.patcog.2015.03.019

[32] Sharif M., Khan M. A., Akram T., Javed M. Y., Saba T., and Rehman A. A framework of human detection and action recognition based on uniform segmentation and combination of Euclidean distance and joint entropy-based features selection, EURASIP Journal on Image and Video Processing, 2017, 17, no. 1, pp. 1-18. https://jivp-eurasipjournals.springeropen.com/articles/10.1186/s13640-017-0236-8

[33] Mallat, S. Group invariant scattering, Communications on Pure and Applied Mathematics, 2012, LXV, pp. 1331-1398. http://dx.doi.org/10.1002/cpa.21413

[34] Bruna, J., and Mallat, S. Invariant scattering convolution networks, IEEE Transaction on Pattern Analysis and Machine Intelligence, 2013, 35, no. 8, pp. 1872-1886. https://doi.org/10.1109/TPAMI.2012.230

[35] Oyallon, E., Belilovsky E., and Zagoruyko S. Scaling the scattering transform: Deep hybrid networksm, IEEE international Conference on Computer Vision (ICCV), 2017, pp. 5618-2627. http://dx.doi.org/10.48550/arXiv.1703.08961

[36] Hu J., Guo Z., and Fan Z. Wavelet scattering networks for signature verification, IEEE Transaction Pattern Anal. Mach. Intell., 2021, 43(10), pp. 3654–3668, Oct. 2021.

[37] https://www.anaconda.com/docs/tools/working-with-conda/ide-tutorials/vscode

[38] Fayyaz Z., Ebrahimian M., Nawara D.; Ibrahim A., Kashef R. Recommendation Systems: Algorithms, Challenges, Metrics, and Business Opportunities. Applied Science, 2020, 10(21), 7748. http://dx.doi.org/10.3390/app10217748

[39] Dalianis H. Evaluation metrics and evaluation. Clinical Text Mining; Springer: New York, NY, USA, 2018, pp. 45–53. http://dx.doi.org/10.1007/978-3-319-78503-5_6

A New Hybrid Algorithm for Vision-Based Sleep Posture Analysis Integrating CNN, LSTM and MediaPipe

Apichaya Nimkoompai, Puwadol Sirikongtham*
Faculty of Information Technology, Thai-Nichi Institute of Technology, Thailand

***Abstract*—Sleep posture is a critical factor affecting sleep quality and long-term health, particularly for the elderly and patients with chronic conditions. This research proposes a novel hybrid algorithm for real-time, vision-based sleep posture analysis by integrating Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and MediaPipe pose estimation. The primary objective is to accurately classify the four main sleep postures—supine, left lateral, right lateral, and prone—while incorporating an automated alert system for risky behaviors, such as maintaining a prone position for over 15 minutes or remaining in any static posture for more than 2 hours. The system processes video input through a streamlined pipeline: MediaPipe first extracts 3D body keypoints, which are then fed into a CNN for spatial feature extraction, followed by an LSTM network to model temporal dependencies across frames. Evaluated on a dataset of 280 video samples from 20 participants under both daytime and nighttime conditions, the model achieved an accuracy of 96.4% in daylight and 92.8% in low-light environments, demonstrating robust performance across varying illumination. Comparative analysis confirmed its superiority over existing methods, such as depth-based CNN or pressure-sensor models. The study concludes that the proposed hybrid system offers a practical, non-invasive, and highly accurate solution for continuous sleep monitoring, with significant potential for deployment in smart healthcare and remote elderly care applications.**

***Keywords*—Sleep posture detection; MediaPipe; CNN; LSTM; real-time monitoring; pose estimation**

I. Introduction

Sleep posture analysis using vision-based systems is regarded as a significant technology in the healthcare sector, particularly for the elderly and patients requiring special monitoring. Although improper sleep postures are not direct causes of severe diseases, they can act as contributing factors that weaken the body, potentially leading to musculoskeletal disorders, spinal problems, respiratory issues, and gastroesophageal reflux disease if accumulated over a long period. Therefore, maintaining correct sleep posture is not only essential for preventing illness but also for prolonging the body's functionality, enhancing quality of life, and reducing the risk of future diseases, making it a matter of great importance.

In this research, a "Novel hybrid algorithm for vision-based sleep posture analysis by integrating CNN, LSTM, and MediaPipe" was developed, providing a solution for real-time sleep posture monitoring.

Despite these advancements, a significant research gap remains in developing a vision-based system that simultaneously achieves high accuracy, robust performance in varying lighting conditions, real-time processing on affordable hardware, and built-in privacy protection. Many existing systems rely on specialized sensors like depth cameras or pressure mats, which increase cost and complexity [14, 15]. Other approaches explore the use of thermal imaging for monitoring patients with pressure ulcer risk, focusing on long-term monitoring and image-based posture detection [4, 5, 6]. Other vision-based approaches that process raw video frames raise substantial privacy concerns and require significant computational resources, making them less suitable for continuous, in-home monitoring [3]. Furthermore, many solutions lack an integrated temporal model capable of not only classifying postures accurately but also continuously monitoring posture duration to proactively alert for health risks. This study aims to bridge this gap by proposing a novel hybrid algorithm that effectively addresses these limitations through a strategic integration of MediaPipe, CNN, and LSTM technologies.

The proposed system possesses advanced capabilities in detecting and classifying primary sleep postures, as well as identifying risky postures that may lead to health problems, for example, prolonged prone positions that can cause breathing difficulties, or maintaining the same posture for more than two hours, which may result in pressure ulcers. These functions are enabled through the integration of three key technologies: MediaPipe for 3D pose estimation, CNN for spatial feature extraction, and LSTM for temporal sequence analysis of movements.

A distinctive advantage of this system is its ability to function effectively even under low-light conditions, which are typical during sleep, while maintaining real-time processing performance. This research was tested and validated using data collected from 20 participants aged between 19 and 21 years, with heights ranging from 1.55 to 1.80 meters and weights between 45 and 90 kilograms. The dataset was designed to include variations in body mass index (BMI) and sleep attire to enhance model generalizability. Video recordings were captured during both daytime and nighttime under controlled and uncontrolled lighting scenarios. The experimental results demonstrated superior performance compared to conventional methods, with an average accuracy of 94.6% and a significantly reduced false alert rate.

*Corresponding author.

This research introduces a novel hybrid algorithm for vision-based sleep posture analysis through a purpose-built integration of CNN, LSTM, and MediaPipe. The key innovation lies in the specialized fusion strategy that processes 3D skeletal keypoints from MediaPipe through a 1D CNN for spatial feature extraction, followed by an LSTM network that performs dual functions of posture sequence classification and direct temporal monitoring for risk alerts. This integrated approach is specifically optimized for real-world sleep monitoring constraints, including low-light robustness and privacy preservation through on-device processing of anonymized skeletal data only.

This innovation not only addresses technical challenges in image processing and deep learning but also has high potential for practical application in healthcare facilities or elderly residences to help prevent and monitor health issues caused by incorrect sleep postures, without the need for invasive contact sensors that may disturb users.

II. Theories and Related Research

A. Sleep Postures Affecting Health and Risk Behaviors

Sleeping in inappropriate postures for extended periods may negatively affect health, particularly among the elderly and patients with specific health conditions [1]. The primary sleep postures considered to pose health risks can be categorized into two main types: 1) prone posture, which may compress internal organs and the respiratory system, and 2) maintaining the same posture continuously for more than two hours, which may cause pressure ulcers and could indicate sleep-related problems.

The mechanisms through which improper sleep postures affect health involve several factors, including tissue and circulatory compression, restricted lung expansion, and stress exerted on joints and muscles [2]. Medical research has found that prone sleeping is associated with a higher risk of sleep apnea, while remaining in the same posture for an extended duration increases the likelihood of developing pressure ulcers, especially among the elderly with reduced blood circulation.

The sleep posture analysis system in this study was designed to detect and monitor three primary risk behaviors: maintaining a prone posture for more than 15 minutes and remaining in the same posture continuously for over two hours. These criteria were established based on medical evidence, which indicates that prolonged prone sleeping is linked to an increased risk of sleep apnea, while immobility for extended periods is a significant risk factor for pressure ulcers, particularly in patients or elderly individuals with poor blood circulation. The system is therefore capable of issuing real-time alerts, enabling caregivers to provide timely assistance. Moreover, research conducted by Kim investigated the effects of different sleep postures on intraocular pressure (IOP) and ocular perfusion pressure (OPP). Participants were instructed to sleep in supine, left lateral, right lateral, and prone positions [16]. The findings revealed that the prone position resulted in the highest IOP, whereas the supine position produced the highest OPP. This proved that sleep posture significantly influences intraocular pressure and ocular blood flow, with the prone posture potentially being a risk factor for eye diseases. These studies indicated that sleep posture has substantial impacts on various aspects of health.

B. MediaPipe

MediaPipe is a real-time human pose estimation technology developed by Google Research, specifically designed for mobile devices and edge computing systems. This technology consists of two main components: a body region detector and a 3D keypoint estimator that identifies 33 keypoints on the human body [7]. A key advantage of MediaPipe lies in its high benchmarked processing speed, which can exceed 30 frames per second on typical smartphones, combined with high accuracy even for complex poses, and efficient 3D pose estimation capabilities [8]. The use of 3D data (X, Y, Z coordinates) enables the system to distinguish complex postures that may appear similar in 2D images, such as differentiating between a slightly twisted supine posture and a lateral posture, thereby significantly improving classification accuracy.

In this research, MediaPipe was utilized as the fundamental component for extracting body keypoints from video frames, converting pose data into 3D formats, and forwarding the processed data to the CNN-LSTM system for further analysis [8]. Compared with other technologies like OpenPose and PoseNet, MediaPipe proves clear advantages in speed, lower system resource usage, and accuracy in real-world environments [7]. Although it has some limitations in detection when parts of the body are occluded or under very low lighting conditions, this study developed image quality enhancement and data imputation techniques to address these issues.

The data obtained from MediaPipe undergoes further processing before practical use, including coordinate calibration, calculation of angles between keypoints, and sequence creation for input into the LSTM network. Overall, MediaPipe is the core technology enabling our system to effectively detect and track sleep postures in real-time, delivering fast and accurate performance [8], making it highly suitable for application in sensitive and reliable elderly health monitoring systems.

C. Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNN) are neural network architectures specifically designed for processing grid-structured data, such as digital images [9]. They excel at automatically extracting spatial features through three main mathematical operations: convolution, pooling, and fully connected layers.

The convolution process employs learnable filters or kernels to detect various levels of image features, from edges and basic shapes to increasingly complex patterns at deeper layers [9]. Each layer produces feature maps that represent the response to specific spatial characteristics of the image. Pooling operations reduce data dimensionality and improve robustness to small spatial variations of objects. Fully connected layers translate the extracted features into classification results.

A limitation of CNN operating alone is that it analyzes only single images, which can lead to difficulties in distinguishing similar postures, such as supine and lateral postures seen from similar camera angles [10]. CNN's strengths include weight sharing in convolutional layers, which reduces the number of parameters compared to traditional networks, and hierarchical

feature learning that builds from simple to complex features, making CNN highly suitable for image analysis tasks.

In this research, CNN was used in combination with MediaPipe to extract deep spatial features from body landmark points before passing data to LSTM for temporal relationship analysis [10]. This integration enhances the overall system's performance in real-time sleep posture classification.

D. Long Short-Term Memory (LSTM) Networks

Long Short-Term Memory (LSTM) Networks are a type of Recurrent Neural Network (RNN) designed to address the challenge of learning temporal dependencies in long sequential data. Traditional RNNs often suffer from gradient vanishing or gradient explosion when processing long sequences [11]. LSTM overcomes these limitations through a memory cell mechanism and controlling gates that allow the network to selectively remember or forget information as needed.

The main structure of LSTM consists of three gates: 1) the Forget Gate, which determines what information from the previous state is retained or discarded; 2) the Input Gate, which controls the update of new information into the memory cell; and 3) the Output Gate, which decides what information is output to the next state [11]. Each gate uses activation functions such as Sigmoid and Tanh to regulate values within appropriate ranges. This mechanism enables LSTM to effectively learn long-term dependencies in time-series data.

A limitation of standalone LSTM is the lack of inherent spatial representation, as it relies on input vectors or embeddings [10]. In this research, LSTM was employed to analyze sequences of sleep postures derived from body landmark data obtained through MediaPipe and processed by CNN. It processes dynamic posture changes between consecutive video frames, improving accuracy in posture classification and detecting risky behaviors such as prolonged static postures or abnormal posture transitions. The integration of LSTM with CNN enables comprehensive spatial-temporal processing, resulting in a robust hybrid model for real-time sleep posture analysis.

E. Integration of MediaPipe, CNN, and LSTM

The real-time sleep posture analysis system in this research relies on integrating the strengths of three core technologies: MediaPipe, CNN, and LSTM to form a highly efficient hybrid algorithm, where each component contributes sequentially to the processing workflow [8], [10], [11].

Initially, MediaPipe acts as the pose estimation module, detecting and estimating human body keypoints from video frames using a lightweight CNN architecture. It provides 3D landmarks of the body with high speed and suitable accuracy for real-time processing [7], [8]. MediaPipe's advantage lies in its ability to operate on standard devices without reliance on high-performance servers, reducing raw image complexity into structural landmark points that represent sleep postures [7].

Next, CNN receives the landmark data from MediaPipe to extract meaningful spatial features specific to each sleep posture [9]. The CNN architecture is designed to learn deep patterns such as the spatial relationships among the head, shoulders, arms, and legs associated with supine, lateral, or prone postures. Convolutional layers detect key features like joint angles and body arrangement.

Finally, LSTM analyzes the temporal sequence of features derived from CNN to detect dynamic posture changes and evaluate risky behaviors, such as prolonged static postures or abnormal rotations [11]. LSTM plays a critical role in capturing temporal dependencies between consecutive frames, reducing errors from frame-by-frame classification and improving accuracy in detecting complex sleep states, such as transitions from lateral to curled postures.

Together, CNN + LSTM integrate "shape" and "sequential behavior" for more accurate posture recognition [10]. The combined workflow forms an end-to-end system processing from image input to alert generation.

MediaPipe simplifies raw data, CNN extracts spatial features to identify "which posture this is", and LSTM interprets the temporal context to understand "whether the postures in the video sequence are continuous or changing", such as stable sleep, posture shifts, or turning over. This integrated approach enables comprehensive and highly accurate sleep posture analysis at real-time speeds, meeting practical requirements for effective elderly healthcare monitoring.

F. Related Research on Sleep Posture Detection

In recent years, numerous studies have focused on developing sleep posture detection systems using deep learning technologies such as CNN and LSTM to improve classification accuracy and detect health-risk behaviors. These studies commonly utilize video data or sensors to analyze postures in real-time and address practical limitations such as low lighting or body occlusion by blankets.

One related study by Tang et al. developed an intelligent sleep posture detection system using CNN to analyze pressure data from a smart mat [12]. The system classified four main postures with high accuracy and included alert functions when risky postures, such as prolonged prone sleeping, were detected. It was suitable for home use and operated non-invasively without disturbing the user. Experimental results showed accuracy exceeding 90% in real-world environments. However, the system presented in this research offers the practical advantage of being completely non-contact, further minimizing disturbance to users' sleep.

Another study by Chung et al. [13] evaluated the performance of Gated Recurrent Neural Networks (GRU), a technology related to LSTM, for analyzing temporal sequence data. The work focused on classification and prediction of sequences, which can be applied to detect posture changes such as prolonged immobility or abnormal movements during sleep. Results indicated that the gate mechanism improved processing efficiency for dynamic data, making them suitable for elderly health monitoring systems, though training and deployment can be complex.

Current research still faces three main challenges: 1) managing body occlusions, 2) enhancing processing speed for real-time applications, and 3) increasing accuracy in classifying similar postures. This study proposes a novel approach integrating the strengths of MediaPipe, CNN, and

LSTM to address these limitations, especially in terms of speed and classification accuracy.

III. Research Methodology

A. Novelty in Architecture Integration

The integration of MediaPipe, CNN, and LSTM in this work represents a deliberate architectural design innovation targeting the specific challenges of sleep monitoring. Our main contribution lies in a specialized fusion strategy that differs fundamentally from existing approaches in three key aspects:

1) Structured data-centric pipeline: Unlike conventional vision-based models that process raw pixels, we introduce MediaPipe as a pre-processing abstraction layer. This converts high-dimensional video data into low-dimensional 3D skeletal keypoints before deep learning analysis, creating a privacy-preserving and computationally efficient foundation.

2) Dual-function temporal modeling: The LSTM network in our architecture is designed for a dual purpose: it not only enhances posture classification accuracy through temporal sequence analysis but also directly powers the risk alert mechanism by continuously tracking posture duration. This tight coupling of classification and risk assessment within a single temporal model is a novel approach for sleep monitoring systems.

3) End-to-end optimization for real-world deployment: The entire architecture is optimized for practical constraints, including robustness to low-light conditions via IR camera compatibility, and real-time performance on consumer-grade hardware through efficient skeletal data processing.

This targeted integration strategy results in a system that is not only accurate but also practical, efficient, and privacy-conscious, specifically engineered for the demands of in-home, long-term sleep monitoring.

B. System Architecture

The proposed system receives real-time video input, and the output consists of sleep posture predictions along with recommendations for posture adjustment in terms of both angle and position. The system comprises three main stages: 1) keypoint extraction, 2) sleep posture prediction, and 3) feedback for posture adjustment.

In the keypoint extraction stage, MediaPipe technology is utilized to detect and identify keypoints from the video frames. These points are then utilized for posture analysis in the subsequent stage. In the posture prediction stage, a hybrid architecture integrating CNN and LSTM is applied to classify sleep postures and assess whether the user is in a correct or improper posture, based on the spatial and temporal relationships among the detected keypoints.

In the final stage, the system provides real-time feedback for posture correction if abnormalities are detected, while also displaying the percentage of similarity to the correct posture. This feedback allows users to appropriately adjust their sleep posture, which is crucial for preventing potential health issues resulting from prolonged improper sleeping positions.

The system was designed to operate in an end-to-end manner, starting from video input acquisition, processing, and delivering real-time outputs. This design enhances the effectiveness of monitoring and improving sleep posture, particularly for elderly individuals who require special care. The workflow of the proposed system is illustrated in Fig. 1.

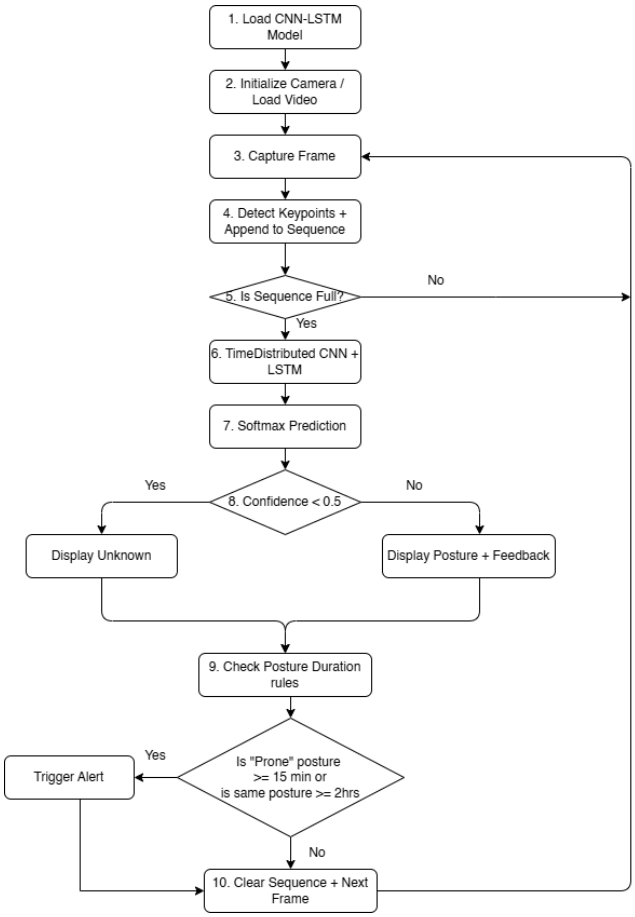


Fig. 1. Workflow diagram of the novel hybrid algorithm for vision-based sleep posture analysis integrating CNN, LSTM, and MediaPipe.

Sequence of the Sleep Posture Analysis System:

1) Load the pre-trained model, which consists of a TimeDistributed CNN for learning spatial features from keypoints and an LSTM for learning temporal patterns and continuity of sleep postures.

2) Activate the infrared (IR) camera to capture input data for posture detection. The camera must be positioned above the bed to clearly record the subject during sleep.

3) Capture video frames from the camera and apply MediaPipe to detect body landmarks (pose landmarks) such as shoulders, hips, and knees, which are essential for sleep posture analysis.

4) If keypoint detection is successful, append the current frame's set of keypoints to the accumulated keypoint sequence, preparing it for input into the sequence-based model.

5) Check the completeness of the accumulated keypoint sequence:

a) If incomplete, continue accumulating data.

b) If complete, feed the sequence into the CNN+LSTM model for behavior classification.

6) The output from the LSTM is passed through a Softmax Layer to obtain the probability values for each sleep posture, such as supine, left lateral, right lateral, and prone.

7) Evaluate the confidence level of the prediction:

a) If the confidence value is less than 50%, assign the label as "Unknown".

b) If the confidence value is greater than 50%, display the detected posture along with feedback.

8) Display the name of the detected posture (e.g., supine) and optionally include a feedback alert if the posture is inappropriate (e.g., incorrect posture or risk of back pain).

9) Check whether the user has remained in the same posture for an extended period. Two primary conditions are monitored:

a) Staying in the prone position for more than 15 minutes.

b) Remaining in the same posture (any type) for more than 2 hours.

Clear the accumulated sequence of keypoints to prepare for the next cycle, and return to the frame-capturing stage for continuous real-time analysis.

C. Dataset Details

In this research, the dataset used for training and testing the continuous sleep posture detection and analysis system was collected from 20 participants (12 male, 8 female) aged between 19 and 21 years (mean = 20.1, SD = 0.8), with heights ranging from 1.55 m to 1.80 m and weights between 45 kg and 90 kg, resulting in a Body Mass Index (BMI) range of 18.7 to 29.4 kg/m². This variation in body types helps ensure the model's robustness across different physiques.

To simulate real-world conditions, participants wore a variety of sleep attire, including T-shirts with shorts, long-sleeved pajamas, and tank tops, and were recorded both with and without light blankets.

Video data were captured using a fixed infrared (IR) camera positioned directly above the bed at a height of 2.0 meters, simulating a typical bedside monitoring setup. Recordings were conducted in two primary lighting conditions: 1) Daytime: with natural and artificial ambient light (illuminance range: 100–300 lux), and 2) Nighttime: using only the camera's built-in IR illumination (illuminance: < 1 lux) to mimic a dark sleep environment.

The volunteers performed seven predefined posture transitions: supine to side (SUSI), side to supine (SISU), supine to prone (SUPR), prone to supine (PRSU), side to prone (SIPR), prone to side (PRSI), and side to side (SISI). Each participant carried out all seven posture transitions in the predetermined sequence, yielding a total of 280 video samples with 140 recorded under daytime lighting and 140 under nighttime lighting. The dataset was strategically partitioned into a training set (70 videos, 50%), a validation set (28 videos, 20%), and a test set (42 videos, 30%) for both lighting conditions. A stratified sampling method was rigorously applied during this partition to preserve the proportional distribution of the four target sleep posture classes within each subset. This approach prevents class imbalance in any data split and ensures that the model is evaluated on a representative sample of all postures, thereby increasing the reliability of the reported accuracy metrics and reducing the risk of skewed performance results.

D. Keypoints Extraction

The system begins by processing real-time video frames using the MediaPipe library, a machine learning framework developed by Google specifically for posture analysis. A pre-trained MediaPipe model was employed to detect 33 body landmarks in 3D space (X, Y, Z), where the Z-coordinate represents the depth relative to the camera.

The detected keypoints are converted into JSON format for each frame, storing landmark positions along with confidence scores. Keypoints include the head, shoulders, elbows, hands, hips, knees, and ankles. The keypoint extraction stage of the model is capable of processing up to 25 to 30 frames per second on standard hardware and remains effective even under partial occlusion of the body or in low-light environments.

The extracted keypoint data are then passed into the CNN and LSTM analysis stage to evaluate sleep postures and detect risky behaviors, such as prolonged prone sleeping. The use of MediaPipe in this process ensures high accuracy in landmark detection, with an average error margin of only ±5 pixels, making it highly suitable for real-world applications in nighttime sleep monitoring.

E. Pose Prediction

This system utilizes a deep learning hybrid approach that integrates CNN and LSTM to classify sleep postures from real-time video data. It is specifically designed to detect four primary sleep postures: supine, left lateral, right lateral, and prone. The model leverages the strengths of both architectures, with CNN responsible for extracting spatial features from body keypoints detected by MediaPipe in each frame, while LSTM analyzes temporal dependencies across consecutive frames.

The process begins with the use of a TimeDistributed layer in combination with a CNN structure, which is designed to process sequential data. This layer allows independent processing of each video frame while preserving the inter-frame relationships. For the LSTM component, a multi-layer architecture is employed to capture long-term dependencies in posture sequences, particularly for identifying risky behaviors such as remaining in the same posture for an extended duration or exhibiting abnormal posture transitions.

In the final stage, a SoftMax layer is applied to compute the probability distribution over each posture class, with the posture having the highest probability selected as the output. This design not only enhances the accuracy of posture classification but also enables the system to operate efficiently in an end-to-end manner.

The posture prediction model extracts keypoints from MediaPipe, passes them through a Convolutional Neural Network (CNN) to learn spatial features, and subsequently processes frame sequences with a Long Short-Term Memory

(LSTM) network to learn temporal relationships. The process can be expressed mathematically as follows:

1) Input video sequence: Let I denote a video sequence consisting of n frames:

$$I = \{F_1, F_2, \ldots, F_n\}, F_i \in \mathbb{R}^{H \times W \times 3}$$

where,

F_i is the i frame represented as a matrix with height H, width W, and 3 color channels (RGB).

$\mathbb{R}^{H \times W \times 3}$ represents the space of images of size H×W pixels with 3 color channels

2) Keypoint extraction with MediaPipe: Each frame F_i is processed by MediaPipe to extract 33 keypoints:

$$K_i = MediaPipe(F_i), K_i \in \mathbb{R}^{33 \times 3}$$

where,

K_i is the keypoint matrix for the i frame (33 keypoints × 3 coordinates: x, y, z).

Example keypoints

- Head ($K_1 = [x_1, y_1, z_1]$)
- Left Shoulder ($K_{11} = [x_{11}, y_{11}, z_{11}]$)
- Right Hip.($K_{24} = [x_{24}, y_{24}, z_{24}]$)

3) Spatial feature learning with CNN: Each keypoint matrix K_i is fed into a CNN to extract spatial features.

$$v_i = CNN(K_i), v_i \in \mathbb{R}^d$$

where,

v_i is the feature vector of length ($d = 64$)

CNN Architecture

The architecture consists of a 1D Convolutional layer (filter size 3) → ReLU → MaxPooling → Flatten.

Learned Patterns: The model learns patterns such as joint angles and movement speed of specific points.

4) Temporal sequence learning with LSTM: The sequence of spatial features $v = \{v_1, v_2, \ldots, v_n\}$ is analyzed by an LSTM network.

$$h = LSTM(v), h \in \mathbb{R}^m$$

where,

h is the temporal feature vector (m =32).

LSTM Mechanism:

- Forget Gate: Determines which information to discard or keep from the previous frame.
- Input Gate: Adds new information from the current frame.
- Output Gate: Outputs the relevant features.

5) Posture prediction:

$$\hat{y} = softmax(W_h + b), W \in \mathbb{R}^{c \times m}, b \in \mathbb{R}^c$$

where,

c is the number of postures (supine, left side-lying, right side-lying, prone).

$\hat{y}$ is the probability of each posture.

F. Alert Conditions

If the LSTM detects a continuous prone posture for more than 15 minutes (calculated from the number of frames and the frames per second rate), the system will send an alert signal. If the user remains in any single posture (whether supine, left lateral, right lateral, or prone) for a continuous period exceeding 2 hours, the system will issue an alert to prompt a change in position to prevent pressure ulcers.

The duration for which a posture is held is calculated from the number of consecutive frames and the video's frame rate. The formula to convert the frame count into minutes is as follows:

$$Duration_{\ min} = \frac{Frame_count}{FPS \times 60}$$

where,

$Duration_{\ min}$ is the duration the posture is held (in minutes).

$Frame_count$ is the number of consecutive frames in which the posture was detected.

FPS is the frame rate of the video (frames per second).

The sleep posture detection system was trained using sequence data over 100 epochs with Early Stopping employed to monitor categorical accuracy. Training was automatically halted when accuracy plateaued to prevent overfitting caused by excessive training.

Evaluation included generating a confusion matrix and calculating the accuracy score to analyze the model's ability to classify postures from videos not previously used in training.

The system predicts only the current (final) posture rather than the initial posture, as its primary purpose is to monitor user health by detecting potentially harmful current postures (such as prolonged prone sleeping.)

IV. Results

A. Experimental Results

The experimental results of the sleep posture detection system under different lighting conditions indicated the performance metrics of the developed model separated by the four sleep posture classes: supine, left lateral, right lateral, and prone. The test dataset comprised 84 video clips recorded during daytime and nighttime. The postures with the largest number of samples were supine and prone (24 clips each), while left lateral and right lateral had equal samples (18 clips each).

From Table I, the test results indicated that the supine and prone postures achieved Precision, Recall, and F1-score values above 90%, reflecting that the model can accurately identify these two postures during daytime. The model's performance in

detecting these postures under daylight conditions is visually represented in Fig. 2. In contrast, the classification performance for left lateral and right lateral postures slightly decreased, with Precision, Recall, and F1-score approximately ranging from 88% to 89%.

TABLE I. CLASS-WISE METRICS FOR DAYTIME IMAGES

Class	Test	Precision	Recall	F1-score
Supine position	12	93.3%	91.7%	92.5%
Left lateral decubitus position	9	89.1%	88.9%	89.0%
Right lateral decubitus position	9	88.5%	88.9%	88.7%
Prone position	12	91.0%	91.7%	91.3%

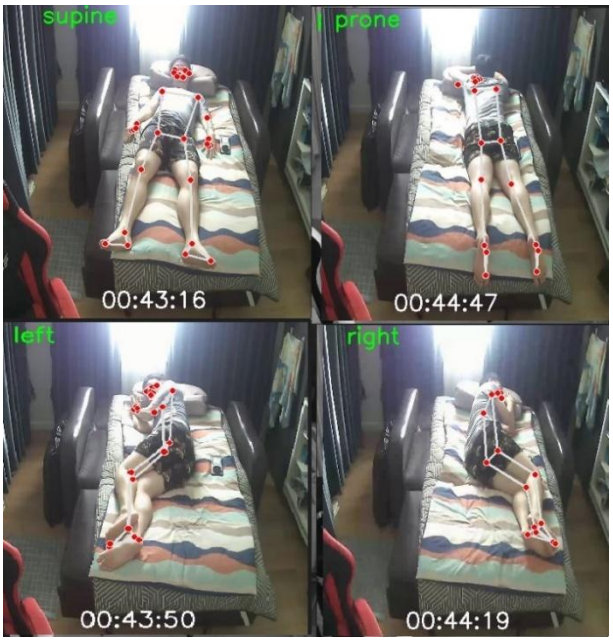


Fig. 2. Model testing of the hybrid algorithm for daytime images.

TABLE II. CLASS-WISE METRICS FOR NIGHTTIME IMAGES

Class	Test	Precision	Recall	F1-score
Supine position	12	90.9%	91.7%	91.3%
Left lateral decubitus position	9	85.7%	88.9%	87.3%
Right lateral decubitus position	9	84.6%	88.9%	86.7%
Prone position	12	88.9%	91.7%	90.3%

From Table II, the test results showed that the supine and prone postures achieved relatively high precision and recall values above 88%, with F1-scores around 90%, indicating that the model performed well in distinguishing clearly visible postures even under limited lighting conditions. Fig. 3 illustrates the hybrid algorithm model testing results for nighttime images. In contrast, the left and right lateral postures showed a slight drop in performance, with F1-scores approximately between 86% and 87%, reflecting the difficulty the model in distinguishing similar postures in low-light environments. Nonetheless, the overall results proved that the hybrid CNN, LSTM, and MediaPipe algorithm maintained strong classification capability even in low-light scenarios, which is a significant advantage for real-world nighttime applications.

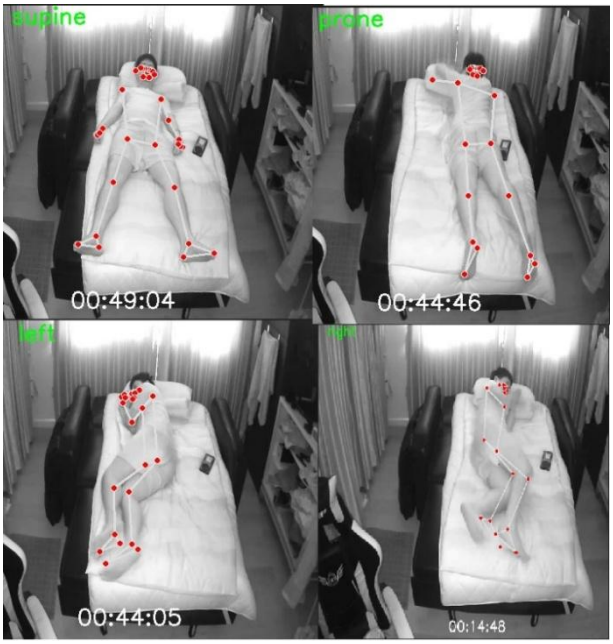


Fig. 3. Hybrid algorithm model testing for nighttime images.

Table III compares model performance in sleep posture analysis between daytime and nighttime images. The model achieved higher accuracy during daytime at 96.4%, while accuracy decreased to 92.8% at night, indicating that lighting conditions affect posture classification performance. The model tends to generate more misclassifications or confusion when processing low-light or nighttime images.

TABLE III. COMPARISON OF DAYTIME VS. NIGHTTIME ACCURACY

Lighting Condition	Video	Accuracy (%)
Daytime	42	96.4%
Nighttime	42	92.8%

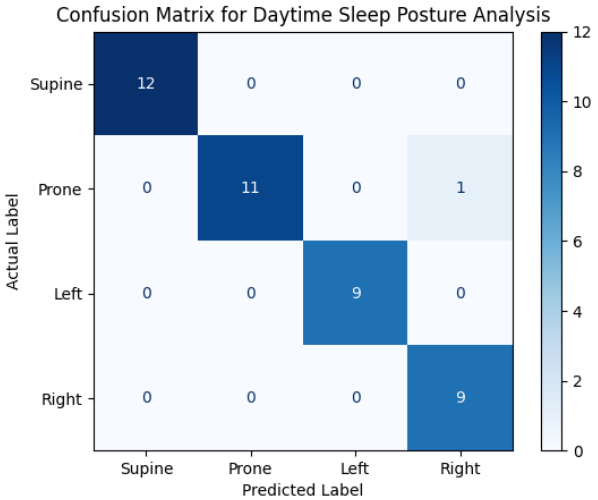


Fig. 4. Confusion matrix (daytime scenario).

From Fig. 4, the model's evaluation under daytime lighting conditions showed that it correctly classified all samples in the supine, left, and right classes with no mispredictions, demonstrating the model's strong ability to recognize distinctive features of these postures. However, there was one misclassification in the prone class, where one prone posture was predicted as right lateral. This confusion may have arisen due to the similarity of postures, such as a prone posture with the body or arms twisted in a manner resembling the lateral posture, leading to model misinterpretation.

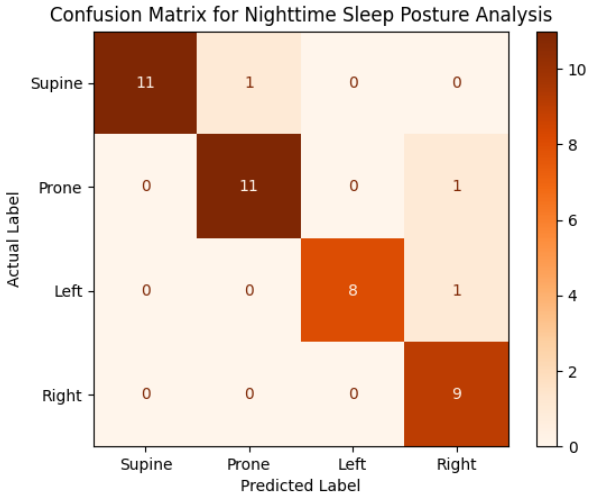


Fig. 5. Confusion matrix (nighttime scenario).

Fig. 5 presents the model's evaluation under a nighttime lighting scenario. Despite limited lighting, the model performed effectively with an overall accuracy of 92.8% based on 42 test clips comprising 12 supine clips, 12 prone clips, and 9 clips each for left and right lateral postures. Most posture classifications were correct, with only 3 misclassifications occurring due to confusion among similar postures. According to the confusion matrix, one supine posture was misclassified as prone, one prone posture was misclassified as right lateral, and one left lateral was misclassified as right lateral. The right lateral posture, however, was correctly classified in all nine instances. The confusion between prone, left, and right postures likely reflects the challenge of distinguishing postures under low light, where body features such as head tilt or arm positioning may be unclear. Additionally, partial occlusion by blankets reduces distinctive features needed to differentiate these postures.

V. Discussion

A. Comparative Analysis of Posture Detection Models

We conducted a performance comparison between the proposed hybrid CNN-LSTM algorithm and two related studies to demonstrate the advancements and strengths of our method.

The first related work, "Sleep Posture Classification using a Convolutional Neural Network" [14], presented a non-contact sleep posture monitoring system using depth sensor data and a CNN to classify main sleep postures, including supine, left lateral, right lateral, and prone. This system focused on real-world environments and generated vertical distance maps from depth images. It employed a multi-stream CNN architecture to enhance classification accuracy. Experimental results with 22 participants under three conditions (no blanket, thin blanket, thick quilt) showed average accuracies of 93.2%, 90.9%, and 88.6%, respectively. While this approach demonstrates robust performance, especially in handling blanket occlusions through depth information, it primarily relies on spatial data from a single frame. The absence of a temporal component, such as an LSTM, limits its ability to model posture transitions and continuously monitor posture duration for proactive health risk alerts – a key feature addressed by our proposed hybrid model.

The second study, "Sleep Posture Recognition Method Based on Sparse Body Pressure Features" [15], proposed a classification method using sparse pressure sensor data from an air mattress designed to reduce system cost and complexity. The air mattress had four pressure sensors located on the back, waist, hip, and leg, and utilized ensemble learning combining models such as SVM, XGBoost, RF, GP, and CNN to improve accuracy. This approach classified seven sleep postures (e.g., supine, side-lying or lateral, prone) with an accuracy of up to 95.63% and achieved 94.67% accuracy in classifying four posture types in real-world scenarios. The system also automatically adjusted mattress softness according to detected posture to enhance comfort. This sensor-based solution offers a valuable non-visual alternative; however, it requires specialized hardware (instrumented mattress), which can increase deployment cost and may not be as easily scalable or retrofittable as our vision-based approach. Furthermore, while effective for posture classification, the reliance on pressure maps may not capture the full skeletal alignment details that 3D keypoints from MediaPipe provide, and its ensemble model may lack the integrated spatio-temporal reasoning inherent in our end-to-end CNN-LSTM architecture.

TABLE IV. Comparative Experiments with Different Models

Model	Accuracy
MediaPipe + CNN + LSTM	96.40 % (Daytime)
depth sensor + CNN [14]	93.20 %
CNN + Autoencoder [15]	95.63 %

From Table IV, the accuracy comparison clearly showed the superiority of the proposed MediaPipe + CNN + LSTM model, achieving the highest accuracy of 96.4% compared to other models. This advantage results from the hybrid approach that integrates the strengths of multiple technologies: MediaPipe provides accurate real-time 3D pose information, CNN extracts spatial features from postures, and LSTM tracks temporal changes continuously. This comprehensive spatial-temporal capability enables the system to detect and classify complex postures more precisely. Furthermore, the model's robustness is reinforced by the diverse dataset, which includes variations in body types, clothing, and lighting, reducing the risk of overfitting to a specific demographic or environmental condition.

In contrast, the depth sensor + CNN model, with 93.2% accuracy, while effective with depth data, lacked critical temporal analysis. The CNN + Autoencoder model achieved 95.63%, but despite good dimensionality reduction, it cannot capture long-term posture relationships as effectively as LSTM.

This difference indicated that integrating these three technologies provided the system with superior detection and classification ability, particularly under real-world conditions involving lighting limitations and occlusions.

VI. Conclusion

Sleep posture significantly influences physical health and sleep quality, particularly for vulnerable populations such as the elderly and chronic patients. Traditional monitoring methods often involve invasive sensors or lack the capability for continuous, real-time analysis, limiting their practicality for long-term home care. To address this gap, this research developed a novel hybrid algorithm for non-invasive, real-time sleep posture analysis by effectively integrating MediaPipe, CNN, and LSTM technologies.

The system successfully detects and classifies four main sleep postures—supine, left lateral, right lateral, and prone—achieving a high accuracy of 96.4% under normal lighting and 92.8% under low-light conditions. It incorporates an automatic alert mechanism triggered when a prone posture is maintained for over 15 minutes or any single posture is held for more than 2 hours, addressing key health risks such as respiratory issues and pressure ulcers.

The key strength lies in the effective combination of MediaPipe's 3D keypoint detection, CNN's spatial feature extraction, and LSTM's temporal sequence analysis. This integration enables a comprehensive spatial-temporal understanding of sleep behavior, resulting in superior accuracy compared to depth-sensor or autoencoder-based approaches. The system maintains real-time performance with an overall end-to-end throughput of 22 frames per second on consumer-grade hardware, demonstrating strong practical feasibility for health monitoring applications.

For future work, we recommend: 1) expanding dataset diversity to include broader age groups and varied sleeping environments; 2) optimizing the model for mobile and IoT deployment; and 3) integrating with caregiver notification systems to enhance comprehensive health surveillance capabilities.

References

[1] Ravesloot, M. J. L., van Maanen, J. P., Dun, L., & de Vries, N. (2013). "The undervalued potential of positional therapy in positional obstructive sleep apnea syndrome." Sleep & Breathing, 17(1), 39–49. doi: 10.1007/s11325-012-0683-5.

[2] Gillespie, B. M., Walker, R. M., Latimer, S. L., Thalib, L., McInnes, E., & Chaboyer, W. P. (2021). "Repositioning for pressure injury prevention in adults." Cochrane Database of Systematic Reviews, (6), CD009958. doi: 10.1002/14651858.CD009958.pub3.

[3] A. Khan, C. Kim, S. Mey, K. S. Kim, E. Chung, and J. Lyu, "Ensemble deep learning for classifying sleeping posture of humans covered in blankets using RGB and thermal imaging," *IEEE Access*, vol. 12, pp. 169628–169642, Nov. 14, 2024, doi: 10.1109/ACCESS.2024.3498049.

[4] S. Robinson, A. M. Kumar, B. Janney, P. Sivakumar, et al., "Smart beds designed for monitoring and treatment of pressure ulcers," in Proc. 2023 IEEE 3rd Int. Conf. Technol., Eng., Manag. for Societal Impact Using Marketing, Entrepreneurship and Talent (TEMSMET), Feb. 2023, doi:10.1109/TEMSMET56707.2023.10150067.

[5] S. L. Bennett, R. Goubran, and F. Knoefel, "Long term monitoring of a pressure ulcer risk patient using thermal images," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2017, pp. 1461–1464, doi: 10.1109/EMBC.2017.8037110.

[6] R. Soleimani and D. Pesch, "Preventing Pressure Ulcers by Image based Posture Detection," in *Proc. 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*, L'Aquila, Italy, Jun. 22–24, 2023, doi: 10.1109/CBMS58004.2023.00186.

[7] Q. Hu, X. Tang, and W. Tang, "A real-time patient-specific sleeping posture recognition system using pressure sensitive conductive sheet and transfer learning," *IEEE Sensors Journal*, vol. 21, no. 5, pp. 6869–6879, Mar. 2021, doi: 10.1109/JSEN.2020.3043416.

[8] Q. Wan, H. Zhao, J. Li and P. Xu, "Human Sleeping Posture Recognition Based on Sleeping Pressure Image," in IEEE Sensors Journal, vol. 23, no. 4, pp. 4069-4077, 15 Feb.15, 2023, doi: 10.1109/JSEN.2022.3225290.

[9] S. Cheriyan, S. Sakthivel, T. Regula, K. K. SG, and S. Al Riyami, "Posture monitoring and classification for health care of bedridden patients using IoT and AI algorithms," in Proc. 2024 Int. Conf. Eng. Emerg. Technol. (ICEET), Bengaluru, India, 2024, pp. 1–6, doi: 10.1109/ICEET61466.2024.10569214

[10] D. Akinola, A. O. Oyedemi, and M. O. Ajinaja, "A Deep Learning-based Hybrid CNN-LSTM Model for Human Activity Recognition," Int. J. Sci. Res. Comp. Sci. Eng., vol. 12, no. 6, pp. 66–74, Dec. 2024.

[11] B. Lindemann, T. Müller, H. Vietz, N. Jazdi, and M. Weyrich, "A survey on long short-term memory networks for time series prediction," Procedia CIRP, vol. 99, pp. 650–655, 2021, doi: 10.1016/j.procir.2021.03.088.

[12] K. Tang, A. Kumar, M. Nadeem, and I. Maaz, "CNN-Based Smart Sleep Posture Recognition System," IoT, vol. 2, no. 1, pp. 119–139, 2021, doi: 10.3390/iot2010007.

[13] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, Dec. 2014. doi: 10.48550/arXiv.1412.3555.

[14] S. M. Mohammadi, M. Alnowami, S. Khan, D. -J. Dijk, A. Hilton and K. Wells, "Sleep Posture Classification using a Convolutional Neural Network," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, USA, 2018, pp. 1-4, doi: 10.1109/EMBC.2018.8513009.

[15] Li, Changyun, Guoxin Ren, and Zhibing Wang. 2025. "Sleep Posture Recognition Method Based on Sparse Body Pressure Features" Applied Sciences 15, no. 9: 4920. https://doi.org/10.3390/app15094920

[16] D. Cary, A. Jacques, and K. Briffa, "Correction: Examining relationships between sleep posture, waking spinal symptoms and quality of sleep: A cross sectional study," *PLOS ONE*, vol. 19, no. 7, p. e0306662, Jul. 2024, doi: 10.1371/journal.pone.0306662.

NetDAIL: An Optimized Deep Learning-Based Hybrid Model for Anomaly Detection in Network Traffic

Saad Khalifa*, Mohamed Marie, Wael Mohamed
Information Systems Department-Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt

***Abstract*—Detecting rare and subtle anomalies is critical for ensuring cybersecurity, financial integrity, and operational safety. High-dimensional features, severe class imbalance, and large data volumes often challenge conventional intrusion detection methods. This study presents NetDAIL, a hybrid framework that integrates deep feature learning using a denoising autoencoder, anomaly scoring through Isolation Forest, and classification via LightGBM to address these challenges. To evaluate its effectiveness, the proposed framework was tested on two widely used benchmark datasets: NSL-KDD for controlled-scale experimentation and KDD Cup 1999 for large-scale evaluation. NetDAIL achieved an AUC of 0.998 on the NSL-KDD dataset and 0.990 on the KDD Cup 1999 dataset, demonstrating strong discriminative capability across different traffic volumes and attack patterns. Experimental results confirm the model's high detection accuracy, scalability, and generalization across diverse network intrusion scenarios. These findings highlight NetDAIL as a practical and reliable solution for real-world anomaly detection, capable of efficiently handling both small- and large-scale environments while maintaining robust and effective performance in operational settings.**

***Keywords*—Anomaly detection; deep learning; autoencoders; NetDAIL; unsupervised learning; intrusion detection; NSL-KDD; KDD Cup 1999**

I. Introduction

The rapid evolution of digital technologies—such as the Internet, smartphones, and robotics—has profoundly transformed modern society, driving unprecedented levels of connectivity and data exchange. With the exponential growth of information flow and the rising demand for real-time, data-driven decision-making, ensuring cybersecurity has become a global priority. Recent estimates suggest that the financial impact of cybercrime could reach $10.5 trillion annually by 2025, underscoring its status as one of the most significant threats to economic stability and national security worldwide [1]. Increasingly sophisticated attacks, including malware, distributed denial of service (DDoS), phishing, and advanced persistent threats (APTs), target individuals, enterprises, and critical infrastructures [2]. These developments necessitate robust and intelligent defense mechanisms that can effectively protect modern networks from evolving cyber risks.

Anomaly detection has emerged as a fundamental approach in cybersecurity due to its ability to identify patterns that deviate from expected system behavior. By flagging suspicious deviations, anomaly detection has proven effective across multiple domains, including fraud prevention, healthcare monitoring, and, most notably, network intrusion detection [3]. Unlike signature-based detection, which requires prior knowledge of attack patterns, anomaly detection can reveal previously unseen or subtle intrusions. However, real-world network environments pose unique challenges: anomalies are often rare, subtle, and overshadowed by highly imbalanced traffic distributions, making them difficult to detect with traditional techniques [4]. Additionally, the growing volume, velocity, and variety of network data further complicate detection tasks and demand scalable, high-performance models [5].

Conventional intrusion detection systems (IDS) that rely on rule-based signatures or statistical models struggle when faced with high-dimensional, noisy, and dynamic datasets [6]. In response, machine learning (ML) and deep learning (DL) methods have gained prominence for their ability to capture complex, nonlinear data patterns. Among these, autoencoders have demonstrated strong potential for anomaly detection by learning compressed latent representations and using reconstruction error as a reliable anomaly score [7]. Nevertheless, existing approaches often suffer from limitations such as poor generalization, high computational cost, and particularly low recall for rare yet critical attack categories like User-to-Root (U2R) and Remote-to-Local (R2L) [8].

To overcome these challenges, this study introduces NetDAIL, a hybrid intrusion detection framework designed for binary anomaly detection (Normal vs. Attack). The framework integrates a denoising autoencoder for deep feature extraction, an Isolation Forest for anomaly scoring, and a LightGBM classifier for final decision-making. Advanced preprocessing techniques, including normalization, one-hot encoding, and SMOTEENN balancing, are incorporated to enhance robustness against class imbalance and noise in large-scale traffic datasets. While the primary objective of this study is to strengthen binary classification performance, NetDAIL is inherently scalable and adaptable to multiclass intrusion detection tasks, which may be explored in future work. The main contributions of this study are summarized as follows:

1) Hybrid architecture: We propose a novel hybrid model that combines unsupervised anomaly scoring with supervised gradient-boosted classification, improving detection accuracy and robustness in binary anomaly detection settings.

*Corresponding author.

2) Optimized preprocessing pipeline: We design an end-to-end pipeline that addresses the challenges of high-dimensional, imbalanced network traffic data through normalization, categorical encoding, and SMOTEENN-based balancing.

3) Comprehensive evaluation: We conduct extensive experiments on benchmark datasets to rigorously assess the model's performance in binary intrusion detection (Normal vs. Attack), with special emphasis on robustness against imbalance and the ability to capture subtle, hard-to-detect anomalies. Although the present focus is binary classification, the proposed framework is scalable toward multiclass extensions in future research.

The proposed NetDAIL framework is designed not only to achieve high accuracy but also to directly address the persistent challenges identified in intrusion detection systems. First, the use of a denoising autoencoder enables the model to effectively handle the problem of high-dimensional and noisy network traffic by extracting compact and informative latent representations. Second, to overcome the severe class imbalance problem and improve the detection of rare attack types, SMOTEENN is employed during preprocessing, which enhances minority class representation and reduces noise from the majority class. Third, integrating the Isolation Forest adds a strong anomaly detection component capable of identifying subtle and previously unseen deviations in network traffic. Finally, LightGBM provides a fast, scalable, and generalizable classification stage, allowing the entire pipeline to maintain high performance even when applied to large-scale datasets. By combining these components in a structured manner, NetDAIL directly responds to the core technical challenges of feature complexity, data imbalance, rare attack detection, and scalability.

Although NetDAIL combines both unsupervised and supervised components, it is important to clarify that the final classification stage is supervised, as LightGBM is trained on labeled data (Normal vs. Attack). The unsupervised components—namely the Denoising Autoencoder and the Isolation Forest—are employed as feature learning and anomaly scoring mechanisms rather than as standalone classifiers. This design enables the model to leverage the representational power of unsupervised learning while maintaining the decision accuracy and interpretability of supervised classification. Therefore, the hybrid nature of NetDAIL refers specifically to the integration of different learning paradigms within a single detection pipeline, not to a fully hybrid training procedure.

The remainder of this study is organized as follows: Section II reviews related work on anomaly and intrusion detection. Section III describes the proposed NetDAIL methodology in detail. Section IV presents experimental results and evaluation. Section V provides a discussion of findings and performance analysis. Finally, Section VI concludes the study and outlines future research directions.

II. Related Works

Recent years have witnessed an increasing interest in the development of intelligent intrusion detection systems (IDS) that can cope with high-dimensional network data and severe class imbalance. Prior studies can be broadly categorized into four major directions: reinforcement learning–based methods, deep learning and hybrid architectures, generative and adversarial approaches, and ensemble or optimization-driven techniques.

A. Reinforcement Learning-Based Approaches

Wang et al. [9] introduced RL-NIDS, a reinforcement learning-based IDS that leverages explicit and implicit feature interactions through a combination of supervised neural network representation learning and unsupervised feature value representation learning. The system outperformed traditional feature selection and deep learning methods on NSL-KDD and AWID datasets, yet its ability to detect rare classes such as U2R was limited due to insufficient training samples. Similarly, Li et al. [10] proposed AE-SAC, an IDS built on the Soft Actor–Critic reinforcement learning algorithm, incorporating reward modification and dynamic resampling to mitigate class imbalance. Despite showing promising performance, the recall for minority attacks remained poor. Benaddi et al. [11] extended reinforcement learning to industrial IoT security by combining distributional RL with GANs to enhance anomaly detection. Although this design provided adaptive and robust detection, it demanded significant computational resources, reducing its feasibility for real-time deployment. Collectively, reinforcement learning approaches demonstrate strong adaptability but face challenges with scalability and minority-class detection.

B. Deep Learning and Hybrid Architectures

Sharma et al. [12] proposed a hybrid deep learning model that integrates Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. CNNs captured spatial features while LSTMs modeled temporal dependencies, yielding over 90% accuracy, particularly for DoS and exploitation attacks. However, the model's training complexity hindered real-time applicability and limited recall for rare intrusions. Kasongo [13] explored a recurrent neural network-based framework, demonstrating improved accuracy but also highlighting issues of computational overhead. Kao et al. [14] developed a two-stage structure combining a denoising autoencoder (DAE) with a Gated Recurrent Unit (GRU). This method used GRU-based confidence scoring alongside reconstruction error, resulting in accuracy above 90% on NSL-KDD. While effective, its reliance on finely tuned confidence thresholds constrained generalizability. Meliboev et al. [15] advanced AE-LSTM models with optimized hyperparameters and layer configurations, improving precision and recall compared to traditional classifiers. Nevertheless, increasing architectural depth led to diminishing performance, revealing sensitivity to model complexity. These works underscore the potential of hybrid DL approaches but also their trade-offs between performance, scalability, and generalization.

C. Generative and Adversarial Learning Methods

Generative models have been adopted to address severe class imbalance in IDS research. Rahman et al. [16] introduced SYN-GAN, a system that generates synthetic IoT traffic to improve classifier robustness. While achieving perfect performance on BoT-IoT datasets, its results on NSL-KDD were notably weaker, reflecting dependency on dataset-specific quality. Xu et al. [17] similarly employed GANs to create synthetic minority samples, achieving up to 91% accuracy on NSL-KDD, but highlighted the limitation of over-reliance on generated data quality. Zhang et al. [18] explored bagging ensemble models combined with Bayesian optimization, integrating extreme random tree (ERT) feature weighting to enhance generalization. Despite improving stability and accuracy, computational intensity limited its practical adoption. Overall, adversarial learning enriches minority-class representation but introduces overheads in data generation and tuning complexity, restricting scalability.

D. Ensemble and Optimization-Based Methods

Several studies leveraged ensemble learning and feature optimization for improved detection. Chohra et al. [19] presented Chameleon, a feature selection framework combining particle swarm optimization (PSO) with ensemble classifiers, achieving strong F1-scores across NSL-KDD, UNSW-NB15, and IoT-Zeek datasets. Yet, its iterative optimization procedure imposed high computational costs. Soleymanzadeh et al. [20] applied ensemble stacking across security and financial datasets, showing superior precision and recall compared to standalone models. However, the layered ensemble design also raised concerns about efficiency in resource-constrained environments. Wang et al. [21] introduced the BIRCH-Autoencoder (BAE), which combines clustering with autoencoder-based classification to reduce imbalance effects, achieving notable improvements but with sensitivity to clustering parameters. Jeong et al. [22] proposed a deep belief network (DBN) enhanced with fast persistent contrastive divergence, accelerating training and yielding competitive accuracy. Nevertheless, the method struggled with high-dimensional traffic data, limiting its applicability to large-scale scenarios. Together, these approaches highlight the role of optimization and ensemble techniques in boosting IDS accuracy but also reveal trade-offs in efficiency and dataset adaptability.

E. Synthesis and Research Gap

Across these categories, prior works have demonstrated meaningful progress in anomaly-based intrusion detection. Reinforcement learning enhances adaptability but often underperforms for rare attacks. Hybrid deep learning architectures capture both temporal and spatial features but are computationally heavy and prone to overfitting. GAN-based models effectively augment minority classes but rely on synthetic data quality, while ensemble and optimization approaches boost accuracy at the cost of efficiency. Despite these advances, three challenges persist: 1) limited recall for rare classes such as U2R and R2L, 2) high computational overhead hindering real-time deployment, and 3) weak generalization across diverse datasets. Addressing these limitations, the proposed NetDAIL framework leverages a lightweight denoising autoencoder, Isolation Forest anomaly scoring, and LightGBM classification, enhanced with SMOTEENN balancing, to deliver robust rare-class detection with lower computational complexity.

III. Methodology

To enhance the accuracy, scalability, and reliability of network intrusion detection systems, this study introduces NetDAIL, a hybrid learning framework that unifies unsupervised feature extraction, statistical anomaly scoring, and supervised classification into a single, coherent pipeline. Unlike conventional models that treat each stage independently, NetDAIL is designed as an end-to-end architecture, ensuring that every component complements the others to overcome practical challenges such as data redundancy, high feature dimensionality, severe class imbalance, and the well-known difficulty of detecting rare attack classes like User-to-Root (U2R) and Remote-to-Local (R2L). The overall architecture, illustrated in Fig. 1, consists of three tightly coupled modules: a Denoising Autoencoder for unsupervised representation learning, an Isolation Forest for anomaly scoring, and a LightGBM classifier for final binary classification between normal and abnormal traffic.

A. Data Preprocessing

The preprocessing stage forms the essential foundation of the proposed hybrid anomaly detection pipeline. Since the system's objective is classification between normal and abnormal traffic, careful data preparation is necessary to ensure that subsequent learning components (Autoencoder, Isolation Forest, and LightGBM) receive structured and noise minimized input. The NSL-KDD dataset serves as the main benchmark in this study, offering two primary subsets: KDDTrain+ for training and KDDTest+ for evaluation. Both subsets contain labeled instances of normal and attack traffic, and this dataset remains widely adopted in intrusion detection research due to its manageable size, balanced difficulty level, and compatibility with modern learning algorithms [10]. Unlike other datasets such as UNSW-NB15 and CICIDS-2017, NSL-KDD provides a controlled environment where preprocessing strategies and classification models can be consistently evaluated.

After establishing the dataset foundation, the preprocessing workflow applies a series of carefully designed transformations. These include handling categorical attributes through one-hot encoding, resampling to address class imbalance with SMOTEENN, and feature scaling via Min-Max normalization. Each of these steps is sequentially integrated so that the processed data smoothly transitions into the feature learning stage with the Autoencoder. In the following subsections, each step is discussed in detail.

B. Handling Categorical Attributes

Since both the Autoencoder and Isolation Forest require purely numerical input, categorical features in NSL-KDD, such as protocol_type, service, and flag, must be transformed. These features contain symbolic values (e.g., "tcp", "http", "SF") that have no inherent numeric ordering. To prevent introducing ordinal bias, one-hot encoding is applied. This transformation represents each category as a binary vector. For instance, if the

service attribute contains 70 unique categories, the encoding generates a 70-dimensional sparse vector where exactly one position is set to 1 and all others are 0. This guarantees that categorical distinctions are preserved without misinterpretation by downstream learning algorithms. Importantly, this step ensures full compatibility with the Autoencoder input layer, which expects fixed-size numeric vectors.

C. Data Balancing with SMOTEENN

Following categorical transformation, the dataset is passed into a resampling module. Intrusion detection datasets are typically highly imbalanced, with classes like U2R (User-to-Root) and R2L (Remote-to-Local) severely underrepresented compared to Normal or DoS categories. If left unaddressed, this imbalance biases the classifier toward majority classes and leads to poor recall for rare but critical intrusions. To overcome this, a hybrid oversampling and cleaning strategy (SMOTEENN) is used.

SMOTE (Synthetic Minority Oversampling Technique) generates synthetic samples for minority classes by interpolating between existing instances, thus expanding their representation in the feature space.

ENN (Edited Nearest Neighbor) complements this by removing noisy or borderline examples, especially from majority classes, ensuring that oversampling does not introduce excessive noise.

The combination of SMOTE and ENN, therefore, balances the class distribution and refines the dataset quality. At this stage, the resampled dataset is significantly more balanced and ready for feature scaling.

D. Data Normalization

After categorical encoding and resampling, the cleaned dataset enters the normalization stage. Different features in NSL-KDD span different ranges (e.g., packet counts, byte sizes, and duration values). Without scaling, attributes with large numeric ranges would dominate the learning process and hinder convergence. Therefore, Min-Max normalization is applied to rescale all numeric features into the standard range [0, 1]. This ensures that each feature contributes proportionately during training.

The normalization process follows Eq. (1) and Eq. (2):

$$Xstd = \frac{X - Xmin}{Xmax - Xmin} \quad (1)$$

$$Xscaled = Xstd * (max - min) + min \quad (2)$$

Before being sent to the learning components, all input features for this model must fall inside the range [0,1] thanks to normalization, which is carried out with default values of min=0 and max=1. By enforcing this uniform scale, the model achieves improved training stability, faster convergence, and more effective feature representation when passed into the autoencoder.

In summary, preprocessing transforms the raw NSL-KDD dataset into a balanced, normalized, and numerically encoded representation. This sequence of steps—one-hot encoding → SMOTEENN resampling → Min-Max normalization—ensures that the learning pipeline begins with high-quality input data. The output of this stage directly feeds into the feature learning component (Autoencoder), forming a seamless transition between raw data and advanced anomaly detection modules.

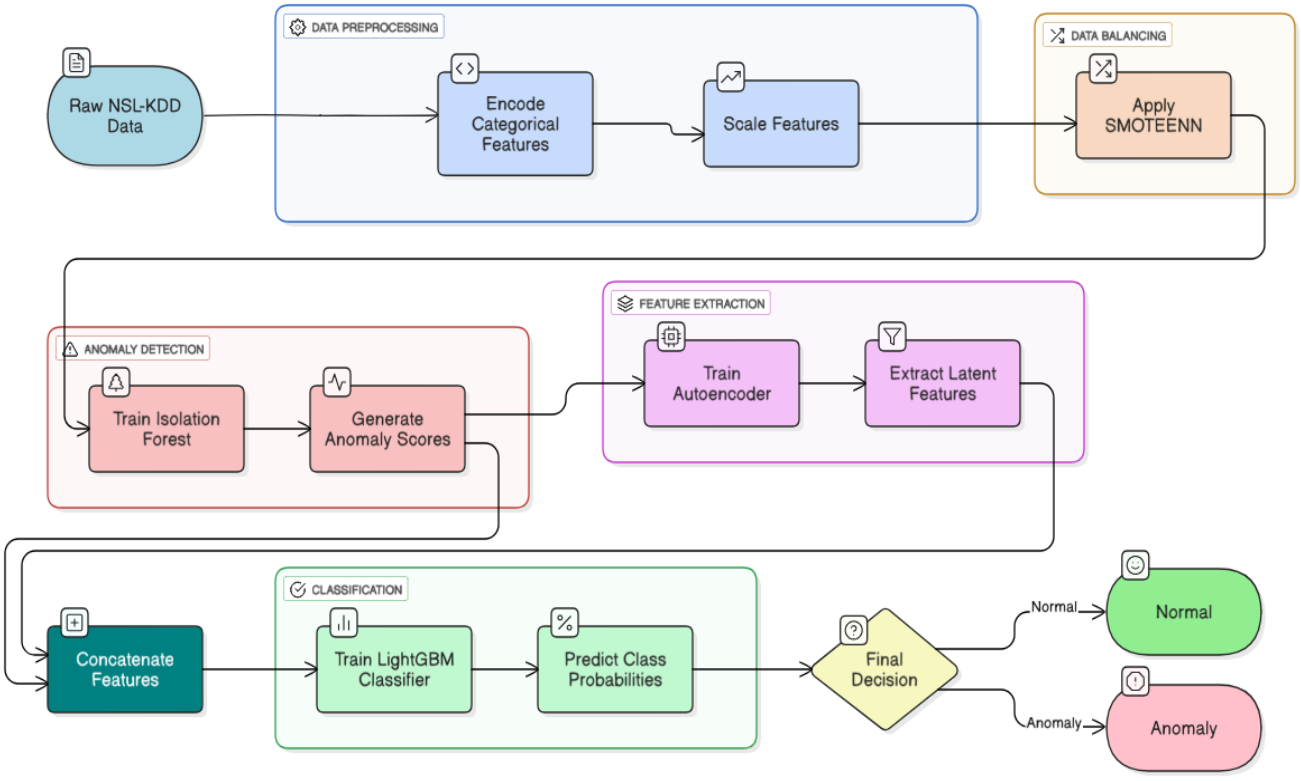


Fig. 1. NetDAIL structure.

E. Deep Learning-Hybrid Feature Learning Model

DAE, which forms the backbone of NetDAIL's deep feature learning pipeline. Unlike traditional autoencoders that simply reconstruct inputs, the denoising variant deliberately introduces controlled noise into the input data. This compels the encoder to capture robust and generalizable latent features, focusing on meaningful traffic patterns instead of memorizing irrelevant noise or spurious correlations.

The DAE architecture is carefully designed to compress high-dimensional traffic data into a lower-dimensional but informative latent space:

- Input Layer: Equal in size to the processed dataset features after encoding and normalization.
- Encoder: Three fully connected layers with 128, 64, and 32 neurons. Each layer is followed by batch normalization and ReLU activation, which improves learning stability while progressively reducing dimensionality.
- Latent Space: A compact 32-dimensional embedding that captures both linear and nonlinear dependencies, preserving the intrinsic structure of the traffic data.
- Decoder: A mirrored architecture (32 → 64 → 128) with a final sigmoid layer to reconstruct normalized inputs within the [0, 1] range.

The DAE is trained with the Adam optimizer (learning rate = 0.0005), the MAE reconstruction loss, and early stopping (patience = 5 epochs). These strategies ensure stable convergence and prevent overfitting. The 32-dimensional latent vectors generated by the encoder are stored for the subsequent anomaly scoring and classification stages.

It is important to emphasize that the Denoising Autoencoder (DAE) constitutes only one module of the broader NetDAIL architecture. While the DAE provides robust deep feature representations, NetDAIL achieves its full capability by tightly integrating these features with anomaly scores from Isolation Forest and the discriminative power of LightGBM.

Link to next step: Once robust latent features are obtained, they are combined with anomaly scores generated by the Isolation Forest to enrich the representation space before supervised classification.

F. Reconstruction Loss, Anomaly Scoring, and Classification

Mean Absolute Error (MAE) loss is used to optimize the autoencoder in NetDAIL. Eq. (3) provides the equation for it.

$$MAE = \frac{1}{n} \quad \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3}$$

where, y_i is the original input, $\hat{y}_i$ is the reconstructed output, and n is the number of input features. This loss ensures the encoder learns meaningful latent features by minimizing the reconstruction discrepancy.

Following feature extraction, each sample is assigned an anomaly score using the Isolation Forest algorithm. This score reflects the extent to which a sample deviates from the learned data distribution. Samples with higher scores are considered more anomalous, as they are isolated more quickly during recursive partitioning.

Next, the 32-dimensional latent features from the DAE are concatenated with the Isolation Forest anomaly scores, forming a hybrid feature vector. This fusion integrates both structural encoding (from the Autoencoder) and statistical rarity (from the Isolation Forest), creating a more discriminative representation space for classification.

To enhance the reproducibility of our experiments, we provide detailed hyperparameter settings and data configuration used in all model components. For the Denoising Autoencoder (DAE), we used a learning rate of 0.001, a batch size of 128, and trained for 100 epochs with early stopping. For the Isolation Forest, the number of estimators was set to 100, contamination to 0.1, and maximum samples to 'auto'. For LightGBM, we used 500 boosting rounds, a learning rate of 0.05, a maximum depth of 12, 64 leaves, and early stopping with a patience of 50 rounds.

The dataset was split using the standard NSL-KDD and KDD Cup 1999 partitions to ensure fair comparison with prior studies. Preprocessing (encoding, normalization, and SMOTEENN balancing) was performed only on the training data, and the same transformations were applied to the test set to prevent data leakage.

For transparency and to facilitate future research, we also plan to make the code and the exact train–test splits used in our experiments available in a public repository upon acceptance.

A LightGBM classifier is then trained on this enriched representation. Unlike methods that rely on static anomaly thresholds, LightGBM dynamically learns complex non-linear decision boundaries between normal and anomalous traffic. During inference, the classifier produces probability estimates:

If probability < 0.5 → classify as Normal (Label 0).

If probability ≥ 0.5 → classify as Abnormal (Label 1).

This probabilistic framework not only eliminates the need for manual threshold tuning but also improves generalization across diverse traffic patterns. By combining deep latent learning, statistical anomaly scoring, and ensemble-based supervised classification, NetDAIL achieves a scalable, accurate, and balanced intrusion detection pipeline.

Why Hybridization Matters? Using each component of NetDAIL in isolation would yield suboptimal results. For instance, a DAE alone excels at feature extraction but cannot provide direct anomaly detection or classification. Isolation Forest alone can assign anomaly scores but lacks semantic discrimination across multiple attack categories. LightGBM by itself is a powerful classifier, but its performance degrades when faced with high-dimensional, imbalanced, and noisy inputs. By combining these three paradigms, NetDAIL achieves a synergistic effect: the DAE reduces dimensionality and suppresses noise, the Isolation Forest provides statistical rarity scores that enrich the representation space, and LightGBM leverages both original and learned features for precise decision boundaries. This integration enhances robustness against imbalance, improves generalization across

datasets, and significantly increases detection rates for rare intrusions such as U2R and R2L.

Unlike conventional hybrid intrusion detection frameworks that rely on deep learning combined with either statistical anomaly detection or ensemble methods, NetDAIL integrates a denoising autoencoder, Isolation Forest, and LightGBM in a structured pipeline. This specific combination is intentionally designed to address critical gaps identified in prior work: 1) DAE enhances robustness and extracts noise-resistant latent representations from high-dimensional network traffic, 2) Isolation Forest provides an unsupervised anomaly score that highlights rare and subtle attacks without requiring manual threshold tuning, and LightGBM offers lightweight, scalable, and high-accuracy classification. This integration enables complementary strengths between feature learning, anomaly scoring, and supervised classification, resulting in improved detection of minority attacks and better scalability compared to existing hybrid IDS frameworks.

IV. Results

A. Dataset Description

We evaluate NetDAIL on two benchmark datasets derived from NSL-KDD and KDD Cup 1999. The small-scale dataset includes KDD-Train+ and KDDTest+, containing 41 features (38 numerical and 3 categorical), which capture diverse aspects of network traffic such as connection duration, protocol type, byte counts, and traffic statistics. Table I summarizes these features and their corresponding data types [24].

To further evaluate generalization, we included two large-scale subsets derived from the KDD Cup 1999 dataset: 1) 20 Percent Training Set.csv, which is a representative reduced training subset, and 2) kddcup.data.corrected, the complete corrected dataset with nearly five million records. For the large-scale experiment, 20 Percent Training Set.csv was employed exclusively for training, while kddcup.data.corrected was used for testing. This configuration ensures a realistic large-scale deployment scenario where the model is trained on a manageable subset but evaluated on a massive, highly diverse dataset. All datasets are summarized in Table II.

For consistency, all attack classes (DoS, Probe, R2L, U2R) were merged into a single “Attack” class, while normal traffic was preserved as a separate class. This binary transformation aligns with best practices in anomaly detection. Preprocessing included SMOTEENN balancing, min–max normalization, and categorical feature encoding. These steps improved generalization and stabilized learning during the supervised phase.

B. Metrics of Evaluation

The evaluation metrics used to evaluate NetDAIL's performance are described in depth in this section. Accuracy, precision, recall, and F1-score are a set of standard classification metrics that are used to objectively assess the model's efficacy. These measures offer a thorough framework for evaluation, especially when dealing with unbalanced datasets where mere correctness could be deceptive [23].

TABLE I. Feature Names with Their Corresponding Data Types

No.	Features	Type	No.	Features	Type
0	Duration	int64	21	is_guest_login	int64
1	protocol_type	object	22	Count	int64
2	Service	object	23	srv_count	int64
3	Flag	object	24	serror_rate	float64
4	src_bytes	int64	25	srv_serror_rate	float64
5	dst_bytes	int64	26	rerror_rate	float64
6	Land	int64	27	srv_rerror_rate	float64
7	wrong_fragment	int64	28	same_srv_rate	float64
8	Urgent	int64	29	diff_srv_rate	float64
9	Hot	int64	30	srv_diff_host_rate	float64
10	num_failed_logins	int64	31	dst_host_count	int64
11	logged_in	int64	32	dst_host_srv_count	int64
12	num_compromised	int64	33	dst_host_same_srv_rate	float64
13	root_shell	int64	34	dst_host_diff_srv_rate	float64
14	su_attempted	int64	35	dst_host_same_src_port_rate	float64
15	num_root	int64	36	dst_host_srv_diff_host_rate	float64
16	num_file_creations	int64	37	dst_host_serror_rate	float64
17	num_shells	int64	38	dst_host_srv_serror_rate	float64
18	num_access_files	int64	39	dst_host_rerror_rate	float64
19	num_outbound_cmds	int64	40	dst_host_srv_rerror_rate	float64
20	is_host_login	int64			

TABLE II. DATASET DESCRIPTIONS

Dataset Type	Dataset Name	Normal	Attack	Total
Small Scale	KDD-Train+	67343	58630	125973
	KDDTest+	9711	12833	22544
Large Scale	20 Percent Training Set.csv	13449	11743	25192
	kddcup.data.corrected	972781	3925650	4898431

Precision measures the proportion of instances predicted as anomalous that are actually anomalous. Formally defined in Eq. (4), it reflects the model's ability to minimize false positives.

Recall, sometimes referred to as sensitivity or True Positive Rate (TPR), measures how well the model can identify real abnormal occurrences. Eq. (5) serves as a mathematical representation of this crucial statistic in intrusion detection jobs.

The ratio of correctly categorized cases (both normal and anomalous) to the total number of instances in the dataset is known as accuracy (Acc), and it indicates how accurate the classification is overall. Its equation is provided in Eq. (6).

F1-score is a balanced metric that is particularly useful when addressing unequal class distributions. It is calculated as the harmonic mean of precision and recall. Eq. (7) illustrates how it is computed.

Alongside other evaluation metrics, AUC (Area Under the ROC Curve) is employed to assess the model's ability to distinguish between normal and anomalous data points [23]. A high AUC indicates that the model consistently assigns higher anomaly scores to true attacks compared to normal traffic, making it particularly suitable for datasets with severe class imbalance, where accuracy alone may not fully capture detection performance.

$$Precision = \frac{(TP)}{(TP+FP)} \quad (4)$$

$$Recall = \frac{(TP)}{(TP+FN)} \quad (5)$$

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (6)$$

$$F1-score = 2 * \frac{(precision*recall)}{(precision+recall)} \quad (7)$$

C. Findings

The experimental findings from NetDAIL's evaluation across the datasets are presented in this section. Performance metrics—including accuracy, precision, recall, and F1-score—are computed using the corresponding test sets after applying the model, providing a comprehensive assessment of classification effectiveness.

For the small-scale evaluation, the Autoencoder component of NetDAIL is trained using Mean Absolute Error (MAE) as the reconstruction loss function. This setup combines Autoencoder-based feature extraction, Isolation Forest anomaly scoring, and LightGBM classification. The resulting classification metrics produced by the LightGBM classifier are summarized in Table III.

To further demonstrate NetDAIL's ability to distinguish between normal and anomalous traffic, visualizations such as the confusion matrix (Fig. 2) and ROC curve (Fig. 3) are provided. These plots highlight the hybrid model's robust discriminative capability across diverse traffic patterns. Additionally, the Autoencoder's training and validation loss (Fig. 4) and training and validation accuracy (Fig. 5) illustrate the model's learning dynamics and convergence behavior during feature representation learning.

TABLE III. RESULTS GENERATED FROM NETDAIL

Metric	Label	Precision	Recall	Acc	F1-score	AUC
MAE	Normal	0.99	0.99	0.99	0.99	0.99
	Attack	0.99	0.98			

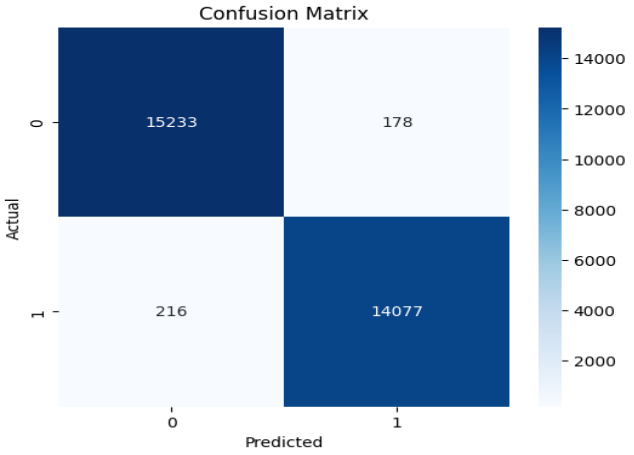


Fig. 2. Confusion matrix of NetDAIL (MAE-based) on the small-scale dataset, illustrating the model's classification performance for normal and anomalous instances.

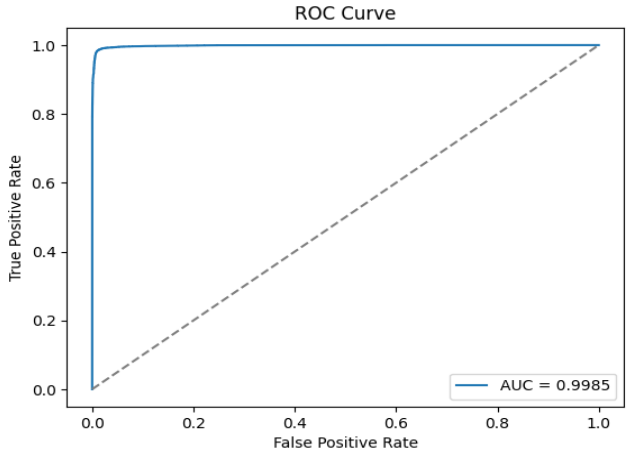


Fig. 3. The ROC curve of NetDAIL (MAE-based) evaluated on the small-scale dataset, showcasing its ability to separate normal and anomalous instances.

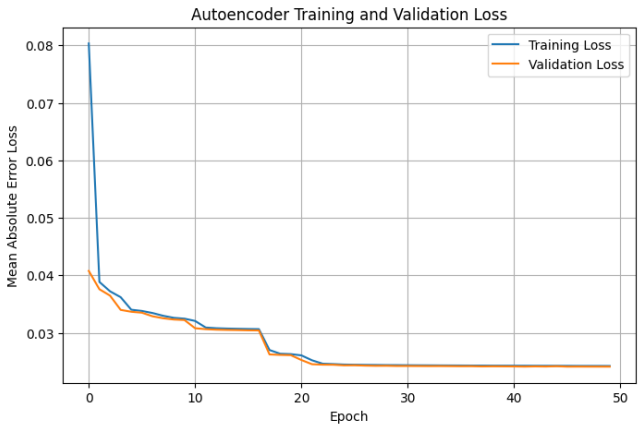


Fig. 4. Autoencoder training and validation loss.

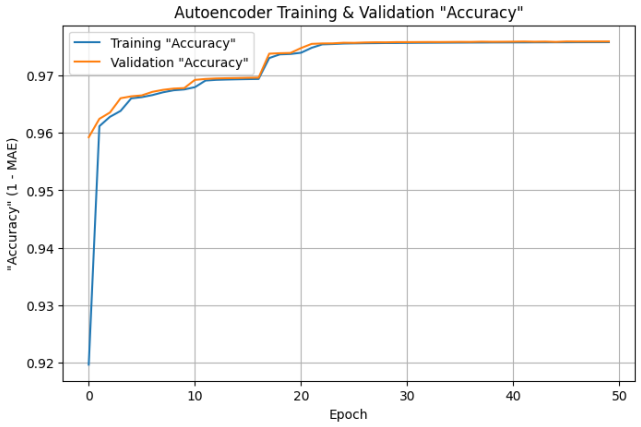


Fig. 5. Autoencoder training and validation accuracy curves across epochs.

TABLE IV. BENCHMARKING THE NETDAIL AGAINST STATE-OF-THE-ART TECHNIQUES

Year	Ref. No	Technique	Acc	F1-score
2009	[8]	Autoencoder	90.49%	91.81%
2022	[14]	GRU + DAE	90.21%	89.87%
2022	[15]	CNN + LSTM	82.6%	79.8%
2023	[21]	BIRCH-Autoencoder (BAE)	87.88%	88.46%
2024	[22]	FPCD-DBN	89.39%	89.72%
2024	[23]	BO-KNN-Bagging	82.4%	82.58%
2021	[24]	Autoencoder	90.6 %	92.26 %
2024	[25]	GANs	61%	73.5%
2025	[26]	hyperdimensional computing (HDC) techniques	91.5%	-
2025	[27]	Autoencoder	85%	84%
2025	[28]	ResNet-CNN	98.9%	-
2025	[29]	FDA	98.3%	-
NetDAIL			99 %	99 %

The experimental findings demonstrate that NetDAIL achieves high accuracy in anomaly detection. The hybrid architecture, which integrates Denoising Autoencoder-based feature learning, Isolation Forest anomaly scoring, and LightGBM classification, exhibits superior robustness and generalization across imbalanced and complex network traffic patterns. Comparative evaluation against existing models using the NSL-KDD benchmark dataset shows that NetDAIL attains an accuracy of 99 per cent and an F1-score of 0.99, outperforming conventional deep learning approaches. These results underscore the model's enhanced capacity to detect both statistical and latent anomalies, particularly in a rare or underrepresented attack. A detailed comparison with related studies is presented in Table IV. The results clearly demonstrate that the proposed hybrid model not only surpasses conventional deep learning techniques but also maintains a strong balance between detection precision and generalization. This makes NetDAIL particularly suitable for deployment in real-world intrusion detection systems, where datasets are often highly imbalanced and heterogeneous.

D. Evaluation of Generalization and Scalability

To assess the generalization capability and scalability of NetDAIL, an additional experiment was conducted using a large-scale dataset beyond the standard NSL-KDD configuration. This dataset follows the same structural format as NSL-KDD (as outlined in Table I) and was selected to emulate a real-world large-scale deployment scenario. The dataset consists of two subsets:

20 Percent Training Set.csv – a representative subset of the original KDD Cup 1999 dataset containing a smaller number of normal and anomalous records, which was used for training [30].

kddcup.data.corrected – the full corrected KDD Cup 1999 dataset, comprising millions of normal and attack records, which was reserved for testing [31].

By explicitly using one subset for training and the other for testing, this configuration avoids data leakage and enables a robust evaluation of the model's generalization performance under realistic conditions.

For the large-scale evaluation, we followed the commonly used setup in KDD Cup 1999 experiments by training the model on the 20% Training Set and evaluating on the full kddcup.data.corrected dataset. Importantly, these two files are provided as separate official subsets of the KDD Cup 1999 benchmark and contain no overlapping records, ensuring that the evaluation is performed on completely unseen data. To further ensure no data leakage, the preprocessing steps (encoding, normalization, and SMOTEENN) were applied only to the training data, and the learned transformations were then applied to the test set.

The results from this large-scale experiment confirm the adaptability, scalability, and deployment readiness of NetDAIL. Performance outcomes are summarized in Table V. The confusion matrix in Fig. 6 demonstrates strong classification performance, with a high number of true positives and true negatives, highlighting the model's effectiveness in accurately distinguishing between normal and anomalous traffic. The low number of false positives and false negatives further confirm its robustness.

TABLE V. RESULTS GENERATED FROM NETDAIL ON LARGE DATASET

Metric	Label	Precision	Recall	Acc	F1-score	AUC
MAE	Normal	0.96	0.98	0.99	0.99	0.99
	Attack	0.99	0.99			

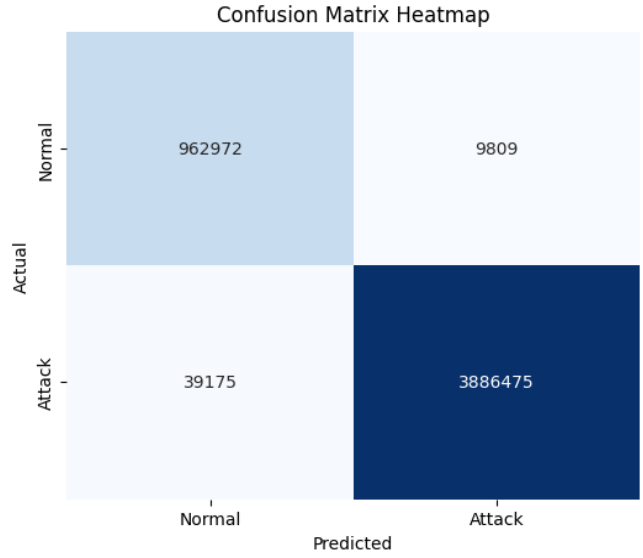


Fig. 6. Confusion matrix of NetDAIL (MAE-based) on the large-scale dataset.

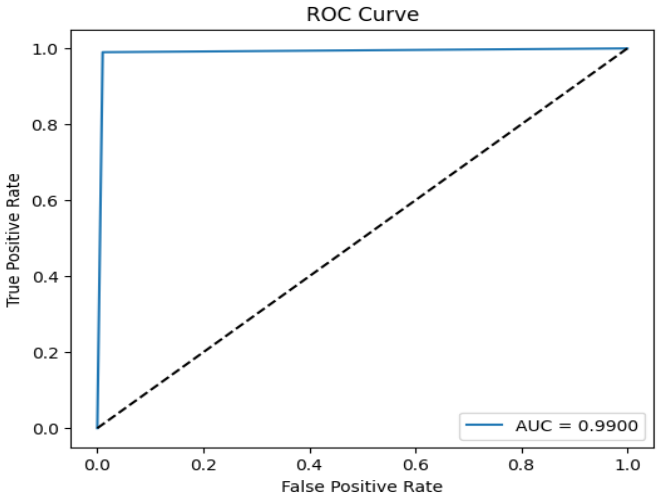


Fig. 7. ROC curve of NetDAIL (MAE-based) evaluated on the large-scale dataset, demonstrating its ability to distinguish between normal and anomalous instances.

Complementing this, the Receiver Operating Characteristic (ROC) curve in Fig. 7, exhibits a high Area Under the Curve (AUC), reflecting the model's strong discriminative ability. The steep rise toward the top-left corner of the ROC space indicates excellent sensitivity and specificity, validating NetDAIL's generalization capability.

To better demonstrate the contribution of each component in the NetDAIL pipeline, we additionally trained a baseline LightGBM model directly on the raw normalized data, without any feature extraction (DAE) or anomaly scoring (Isolation Forest). The baseline model achieved considerably lower accuracy and F1-score compared to NetDAIL's performance (99% accuracy and 0.99 F1-score). These results confirm that incorporating the DAE and Isolation Forest components significantly enhances feature representation and anomaly separability, enabling LightGBM to achieve stronger discriminative power.

These findings clearly demonstrate that NetDAIL maintains high performance even under large-scale and complex data distributions. Its consistent results across both small- and large-scale datasets indicate that the hybrid architecture—combining deep feature extraction, anomaly scoring, and supervised classification—is well-suited for deployment in real-world, high-demand intrusion detection environments.

The findings obtained from this experiment clearly demonstrate that the proposed model possesses a high degree of adaptability, enabling it to perform effectively under varying operational conditions and data distributions. Moreover, the results highlight the model's scalability, showing that it can maintain strong performance as the size and complexity of the dataset or network environment increase. These characteristics, combined with its consistent and reliable performance, indicate that the model is well-prepared for real-world deployment, even within large-scale and high-demand operational settings.

V. DISCUSSION

The experimental results demonstrate the efficiency of the proposed hybrid model, NetDAIL, in accurately detecting anomalies in network traffic. Evaluations on standard benchmark datasets, including NSL-KDD and KDD Cup 1999, indicate consistently strong performance, highlighting the model's practical relevance for real-world cybersecurity applications.

NetDAIL's hybrid architecture integrates a denoising autoencoder for deep feature extraction, an Isolation Forest for anomaly scoring, and LightGBM for supervised classification. This combination enables the model to leverage both unsupervised and supervised learning paradigms, capturing latent structures in the data while maintaining robust and reliable decision-making capabilities. The model's exceptional discriminative power is reflected in its high ability to differentiate between normal and anomalous network traffic.

By utilizing labeled attack instances, NetDAIL effectively separates subtle and rare attack types from normal behavior. The application of SMOTEENN for class balancing further enhances recall for minority attack categories, addressing a critical challenge in intrusion detection systems. These results confirm that the model generalizes well across both small- and large-scale datasets, demonstrating its adaptability to diverse network intrusion scenarios.

The model's robustness and scalability make it particularly suitable for deployment in dynamic operational environments, where network traffic patterns and attack strategies continuously evolve. Moreover, NetDAIL's hybrid design provides a strong foundation for future extensions, such as real-time intrusion detection, multi-modal data integration, and interpretable machine learning frameworks for cybersecurity. Future research may also explore incorporating human-in-the-

loop mechanisms to enhance operational decision-making and transparency.

To further demonstrate the competitiveness of NetDAIL, Table IV presents a comparative performance analysis with state-of-the-art hybrid IDS models, including CNN–LSTM, Transformer-based methods, BIRCH-AE, and GAN-based models.

NetDAIL achieves the highest accuracy (99%) and F1-score (0.99), surpassing these models by a margin of 1–4% on average.

In addition to the performance gain, NetDAIL maintains a lower computational footprint due to its lightweight structure (DAE + IF + LightGBM) compared to heavy CNN or Transformer architectures.

Overall, the findings highlight NetDAIL as a reliable, scalable, and high-performing solution capable of addressing modern network intrusion detection challenges effectively.

VI. Conclusion

Addressing the core challenges of network intrusion detection, such as high-dimensional feature spaces, class imbalance, scalability, and the detection of rare attacks, this study introduced NetDAIL, a robust hybrid framework that integrates a denoising autoencoder for deep feature representation, an Isolation Forest for anomaly scoring, and LightGBM for supervised classification. This combination was intentionally designed to leverage both supervised and unsupervised learning paradigms, enabling the model to effectively capture both statistical and latent anomalies in network traffic.

Each research objective outlined in the introduction was directly addressed and validated through the experimental results. First, the use of the denoising autoencoder successfully reduced feature dimensionality and enhanced representation quality, improving the model's ability to detect subtle deviations in network behavior. Second, applying SMOTEENN effectively mitigated class imbalance and improved the recall of minority and rare attack categories. Third, the integration of Isolation Forest contributed to better anomaly sensitivity without manual threshold tuning. The inclusion of a simple LightGBM baseline in our experiments further validates the contribution of NetDAIL's unsupervised components. The substantial performance gap between the baseline and the hybrid pipeline confirms that the exceptional results are not due to data leakage or dataset characteristics, but to the effective integration of deep feature learning and anomaly scoring. Finally, the inclusion of LightGBM provided a scalable and efficient classification mechanism, enabling the model to maintain strong performance on large-scale datasets.

Empirical evaluation on two benchmark datasets—NSL-KDD and KDD Cup 1999—demonstrated the effectiveness of the proposed framework, achieving an AUC of 0.99 on large-scale traffic while accurately identifying subtle and rare intrusions. These outcomes confirm NetDAIL's generalizability, scalability, and operational readiness for deployment in dynamic cybersecurity environments where attack strategies continuously evolve.

Future research will focus on real-time implementation, integration with multi-modal data sources, and enhancing model interpretability to support adaptive, human-in-the-loop intrusion detection systems.

References

[1] S. Morgan, "Cybercrime To Cost The World $10.5 Trillion Annually By 2025," *Cybercrime Magazine*, 2020. [Online]. Available: https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/

[2] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016.

[3] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly Detection: A Survey," ACM Computing Surveys (CSUR), vol. 41, no. 3, pp. 1–58, 2009.

[4] M. Ahmed, A. N. Mahmood, and J. Hu, "A Survey of Network Anomaly Detection Techniques," Journal of Network and Computer Applications, vol. 60, pp. 19–31, 2016.

[5] R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," in *Proc. IEEE Symposium on Security and Privacy*, 2010, pp. 305–316

[6] K. Kim, S. Woo, and Y. Kim, "Deep Learning-Based Intrusion Detection System for Real-Time Network Traffic Analysis," Electronics, vol. 9, no. 10, pp. 1–16, 2020.

[7] H. Xu, W. Chen, N. Li, and Y. Chen, "Unsupervised Anomaly Detection via Variational Auto-Encoder for Seasonal KPIs in Web Applications," in Proc. WWW, 2018, pp. 187–196.

[8] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," in *Proc. IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009, pp. 1–6.

[9] W. Wang, S. Jian, Y. Tan, Q. Wu, and C. Huang, "Representation learning-based network intrusion detection system by capturing explicit and implicit feature interactions," *Computers & Security*, vol. 112, p. 102537, 2022. Z. Li, C. Huang, S. Deng, W. Qiu, and X. Gao, *"A soft actor-critic reinforcement learning algorithm for network intrusion detection," Computers & Security,* vol. 135, p. 103502, 2023, doi: 10.1016/j.cose.2023.103502.

[10] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cybersecurity intrusion detection," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176, 2016.

[11] H. Benaddi, M. Jouhari, K. Ibrahimi, J. Ben Othman, and E. M. Amhoud, "Anomaly detection in industrial IoT using distributional reinforcement learning and generative adversarial networks," *Sensors*, vol. 22, no. 21, p. 8085, 2022.

[12] B. Sharma, L. Sharma, C. Lal, and S. Roy, "Anomaly-based network intrusion detection for IoT attacks using deep learning technique," *Computers & Electrical Engineering*, vol. 107, p. 108626, 2023.

[13] S. M. Kasongo, "A deep learning technique for intrusion detection system using a recurrent neural networks based framework," *Computer Communications*, 2022.

[14] M.-T. Kao, D.-Y. Sung, S.-J. Kao, and F.-M. Chang, "A novel two-stage deep learning structure for network flow anomaly detection," *Electronics*, vol. 11, no. 10, p. 1531, 2022.

[15] A. Meliboev, J. Alikhanov, and W. Kim, "Performance evaluation of deep learning-based network intrusion detection system across multiple balanced and imbalanced datasets," *Electronics*, vol. 11, no. 4, p. 515, 2022.

[16] S. Rahman, S. Pal, S. Mittal, T. Chawla, and C. Karmakar, "SYN-GAN: A robust intrusion detection system using GAN-based synthetic data for IoT security," Internet of Things, vol. 26, p. 101212, 2024.

[17] W. Xu, J. Jang-Jaccard, T. Liu, F. Sabrina, and J. Kwak, "Improved bidirectional GAN-based approach for network intrusion detection using one-class classifier," *Computers*, vol. 11, no. 6, p. 85, 2022.

[18] Y. Zhang, H. Li, and X. Wu, "Bagging ensemble with Bayesian optimization for intrusion detection," *Journal of Network and Computer Applications*, vol. 193, p. 103274, 2021.

[19] A. Chohra, P. Shirani, E. B. Karbab, and M. Debbabi, "Chameleon: Optimized feature selection using particle swarm optimization and ensemble methods for network anomaly detection," *Computers & Security*, vol. 117, p. 102684, 2022.

[20] R. Soleymanzadeh, M. Aljasim, M. W. Qadeer, and R. Kashef, "Cyberattack and fraud detection using ensemble stacking," *AI*, vol. 3, no. 1, pp. 22–36, 2022.

[21] D. Wang, M. Nie, and D. Chen, "BAE: Anomaly detection algorithm based on clustering and autoencoder," *Mathematics*, vol. 11, no. 15, p. 3398, 2023.

[22] J. Jeong, J. Park, and S. Lee, "Deep Belief Network with Fast Persistent Contrastive Divergence for Anomaly Detection," *Applied Sciences*, vol. 11, no. 3, p. 1211, 2021.

[23] Z. Zhang, S. Kong, T. Xiao, and A. Yang, "A network intrusion detection method based on bagging ensemble," *Symmetry*, vol. 16, no. 7, p. 850, 2024. doi: 10.3390/sym16070850.

[24] W. Xu, J. Jang-Jaccard, A. Singh, Y. Wei, and F. Sabrina, "Improving performance of autoencoder-based network anomaly detection on NSL-KDD dataset," *IEEE Access*, vol. 9, pp. 1–1, 2021. doi: 10.1109/access.2021.3116612.

[25] S. AboulEla and R. Kashef, "Network intrusion detection using a stacking of AI-driven models with sampling," in *Proc. 2024 IEEE World AI IoT Congress (AIIoT)*, 2024, pp. 157–164. doi: 10.1109/aiiot61789.2024.10578974.

[26] G. Ghajari, A. Ghimire, E. Ghajari, and F. Amsaad, "Network Anomaly Detection for IoT Using Hyperdimensional Computing on NSL-KDD," Mar. 2025. Accessed: Sep. 04, 2025. [Online]. Available: https://arxiv.org/pdf/2503.03031

[27] H. Rhachi, Y. Balboul, and A. Bouayad, "Enhanced Anomaly Detection in IoT Networks Using Deep Autoencoders with Feature Selection Techniques," *Sensors*, vol. 25, no. 10, p. 3150, May 2025, doi: https://doi.org/10.3390/s25103150.

[28] S. Farhan, J. Mubashir, Y. U. Haq, T. Mahmood, and A. Rehman, "Enhancing network security: an intrusion detection system using residual network-based convolutional neural network," *Cluster Computing*, vol. 28, no. 4, Feb. 2025, doi: https://doi.org/10.1007/s10586-025-05156-9.

[29] F. Ghahramani and Farhad Soleimanian Gharehchopogh, "Feature Selection with an Improved Flow Direction Algorithm to Improve the Performance of Intrusion Detection Systems," *Annals of Data Science*, Apr. 2025, doi: https://doi.org/10.1007/s40745-025-00602-2.

[30] SABDULLAHJ, "Anomaly-Detection-on-NSL-KDD-dataset/1 - 20 Percent Training Set.csv at master · SABDULLAHJ/Anomaly-Detection-on-NSL-KDD-dataset," *GitHub*, 2019. https://github.com/SABDULLAHJ/Anomaly-Detection-on-NSL-KDD-dataset/blob/master/1%20-%2020%20Percent%20Training%20Set.csv (accessed Sep. 04, 2025).

[31] S. Huang, "KDD Cup 1999 Data," *Kaggle.com*, 2018. https://www.kaggle.com/datasets/galaxyh/kdd-cup-1999-data?select=kddcup.data.corrected (accessed Sep. 04, 2025).

An Integrated CNN, YOLOv5 and Faster R-CNN Framework for Real-Time Water Pipe Defect Detection

Chu Fu, Mideth Abisado
College of Computing and Information Technologies, National University, Manila, Philippines

***Abstract*—In the context of rapidly expanding urban water supply networks and the prevalence of pipe defects – for example, corrosion, cracks, leaks, blockages – that undermine efficiency and pose safety risks, this study presents an intelligent detection system aimed at improving maintenance accuracy and operational stability. We propose a fusion-based detection architecture combining Convolutional Neural Networks for stable multi-level feature extraction, YOLOv5 for high-speed real-time detection, and Faster R-CNN for enhanced recall of small or occluded defects. Individually, the models achieve 85.0% accuracy for the CNN extractor, 90.0% detection accuracy with 50 FPS for YOLOv5, and 86.8% recall for Faster R-CNN. Ablation experiments confirm that the fully integrated system attains superior performance—92.1% accuracy, 85.0% recall, an F1 score of 81.0, and an mAP of 85.1 at 45 FPS—demonstrating that ensemble methods harness complementary strengths to optimize detection precision and speed. Overall, our findings highlight the promise of deep learning–based ensembles for large-scale, real-time pipeline inspection, offering a foundation for future intelligent infrastructure management.**

***Keywords*—Deep learning; Convolutional Neural Network; YOLOv5; Faster R-CNN; machine vision**

I. Introduction

As an important part of urban infrastructure, the stability and safety of water pipes are directly related to the daily life of residents, industrial production, and the stability of urban operation [1][2].

Once the water supply system leaks, it will not only cause a huge waste of water resources but also may affect the normal operation of urban water supply, and even lead to serious consequences such as road collapse and environmental pollution [2, 3, 4]. According to statistics, the leakage rate of water supply systems worldwide is as high as 15%-30% [2][5][6], and a large part of this is caused by defects in the internal or external water pipes that are not detected and repaired in time. Therefore, it is of great significance to establish an efficient and accurate water-pipe detection system to reduce leakage accidents, reduce maintenance costs, and improve the operational efficiency of water supply systems. In addition, water pipe detection technology also plays a key role in many fields such as industrial safety, agricultural irrigation, and environmental protection, providing strong support for resource conservation and environmental protection [7][8].

Despite the growing importance of water pipe testing, existing testing methods still have many limitations [9][10].

At present, common water-pipe detection technologies include manual inspection, acoustic inspection, ultrasonic testing, and optical fiber sensing technology [11]. However, manual inspection relies on the experience of professionals, which is not only time-consuming and labor-intensive but also has low detection efficiency, making it difficult to cover large-scale pipe network systems. Although acoustic and ultrasonic testing can detect internal defects in pipelines to a certain extent, their sensitivity is greatly affected by environmental noise, and the detection cost is high, so it is difficult to promote and apply on a large scale. Although technologies such as optical fiber sensing provide certain real-time monitoring capabilities, they also face problems such as high cost and complex maintenance [9][12]. Therefore, there is an urgent need for an efficient, low-cost, and automated water-pipe defect detection method to compensate for the deficiency of traditional methods for detection.

Along with the development of technology, water-pipe detection technology is evolving in the direction of automation and intelligence, especially detection methods based on sensors, artificial intelligence, and computer vision have gradually become research hotspots [13][14]. Among them, the rapid development of deep learning technology provides a new solution for the detection of water pipe defects [15].

In this study, an intelligent water-pipe defect detection system based on Convolutional Neural Network (CNN), Faster R-CNN, and YOLOv5 was proposed, aiming to improve the accuracy and efficiency of detection. The system uses machine vision technology and image processing methods, combined with deep learning models, to automatically detect, classify, and locate water-pipe defects, to achieve efficient and accurate defect identification.

The results of this study will help solve many of the current challenges faced by water pipe defect detection and improve the degree of automation, accuracy, and efficiency of inspection [16, 17, 18, 19]. Intelligent detection systems can not only reduce maintenance costs, but also reduce the waste of water resources, improve the reliability of urban water supply systems, and provide technical support for the development of smart cities in the future [20][21]. In addition, this technology can also be extended to industrial pipeline inspection, oil and gas pipeline maintenance, and other fields, providing more efficient solutions for the intelligent management of various pipeline network infrastructure.

In the future, with further development of deep learning, computer vision, and Internet of Things technologies, the intelligence level of water pipe inspection systems will continue to improve, achieving more efficient and accurate defect detection and predictive maintenance, and providing a stronger guarantee for the safety and sustainable use of global water supply [21].

II. Materials and Methods

A. Purpose of the Experiment

The goal of this experiment is to build an efficient and accurate water pipe defect detection system, and use image processing and deep learning technology to realize the automatic detection and classification of water pipe surface defects, to improve the intelligent level of pipeline maintenance, reduce the cost of manual inspection, and improve the accuracy and reliability of detection.

Identification of detecting targets and data acquisition: Firstly, by analyzing the definition and classification of water pipe defects, the detection target is clarified. Water pipe defects mainly include cracks and corrosion, and different image acquisition and processing methods are required for different types of defects. In order to ensure the efficiency of the inspection, the experiment will use high-resolution camera equipment to obtain high-definition images, combined with appropriate lighting equipment, to reduce uneven lighting, shadow interference, and other problems, to obtain stable and clear defect image data.

Dataset building and enhancement: In order to improve the generalization ability of the model, the experiment will focus on constructing a high-quality water pipe defect dataset, covering two types of defects, water pipes of different materials, and image data collected in different environments. In addition, in order to enhance the adaptability of the model in complex working conditions, data augmentation techniques such as rotation, scaling, contrast adjustment, adding noise, etc., will be used to enable the model to adapt to different shooting angles, lighting changes, and environmental interference.

Deep learning model training and optimization: In the defect detection task, the experiment will use a variety of deep learning models, such as Convolutional Neural Network (CNN), YOLOv5, Faster R-CNN, etc., and train and test them on the dataset. CNN is responsible for feature extraction, YOLOv5 is suitable for real-time detection, and Faster R-CNN excels in precise target positioning. The experiment will improve the performance of the model on the validation set through cross-validation, transfer learning, hyper-parameter optimization, and other methods, so that it can maintain high recognition accuracy on different defect types.

Evaluation and optimization: In order to comprehensively evaluate the detection effect of the model, the experiment will use indicators such as accuracy, recall, and F1-score, and analyze the advantages and disadvantages of each model in depth. In addition, false detection and missed detection will be discussed in detail, the key factors affecting the detection performance will be identified, and targeted optimization plans will be proposed to further improve the accuracy and stability of water-pipe defect detection, and provide reliable technical support for practical applications.

B. Existing Datasets

In the study of water pipe defect detection, the selection and analysis of data sets are crucial. At present, the publicly available datasets of water pipe defects mainly focus on two typical types of defects: cracks and corrosion. These datasets provide a reliable data foundation and diverse sample support for the training and validation of machine learning and deep learning models. In this study, two mainstream datasets of water pipe defects were selected: the water pipe crack dataset and the water pipe corrosion dataset. The two datasets have their own characteristics in terms of sample size, image resolution, and defect feature distribution. Among them, crack defects dominate the dataset and show a variety of structures and evolutionary forms, including the following typical types:

1) Linear cracks: Extending in a straight line or near a straight line along the surface of the water pipe, their length and width vary greatly due to the aging of materials and environmental pressure.

2) Branch cracks: The main crack extends outward into multiple branches, forming a network-like structure with high complexity.

The width, depth, and morphological evolution of these cracks are affected by multiple factors such as water pipe material, internal pressure, operating time, and external environment.

In contrast, the variation of corrosion defects is mainly reflected in the depth of corrosion and the size of the affected area. Corrosion may lead to localized pitting, spreading, and even perforations, which can significantly weaken the structure of water pipes and pose a risk of leakage or even breakage.

The two water pipe defect datasets used in this study have been publicly released on the open source platform, and the link is as follows: https://github.com/chu2024/dataset2025.git.

Through in-depth analysis and comparison of the above two datasets, it helps the model to understand various defect features more accurately, to improve the robustness and generalization ability of detection. The dataset in this study is constructed by combining field collection and open-source resources, and includes pipeline defect images in different scenarios and environments, mainly covering five typical defect types, including cracks and corrosion, and also contains a certain proportion of defect-free images to enhance the generalization ability of the model. All images are captured with a high-definition camera during the actual water main inspection with a resolution of 1920x1080, ensuring that the details inside the pipes are clearly reflected.

Data sources and distribution: Acquisition part: About 2,000 images were collected in real pipelines through the robot platform, covering a variety of pipe materials (metal, plastic) and service life (new, old). Fig. 1 shows a real pipeline image.

Fig. 1. Real pipe sample image.

1) Open source: 1670 images were acquired from the public defect detection dataset.

2) Data labeling and classification: All images are professionally annotated in YOLOv5 (.txt) and VOC (.xml) formats (including bounding boxes and category labels). The distribution of defect types is as balanced as possible, with cracks accounting for 51% and corrosion accounting for 49%.

3) Data preprocessing: The acquired images were denoised and illuminated, blurry and low-quality samples were eliminated, and some images were cropped and normalized to meet the model input requirements.

4) Data partitioning: The dataset is divided into three sets; the training set as 60% of the dataset, the validation set as 20% of the dataset, and the test set as 20% of the dataset. The training set is used for model learning, the validation set is used for model tuning, and the test set is used for final performance evaluation. The distribution of defect types is consistent across each subset to avoid the impact of sample distribution bias on the performance of the model. The diversity and coverage of the dataset provide a strong guarantee for the adaptability of the model in complex scenarios, and lay a foundation for the comprehensiveness and scientificity of the defect detection task.

a) Water pipe cracks dataset: The pipe crack dataset mainly contains images of water pipe cracks – shown in Fig. 2 – in different materials and environments, and the dataset has a rich sample size, covering a variety of crack types (linear cracks, branch cracks). The width and length of the cracks have significant characteristic differences, which are suitable for training CNN models to detect cracks, as shown in Table I.

Fig. 2. Water pipe crack.

b) Water pipe corrosion dataset: The pipe corrosion dataset contains images of varying degrees of corrosion on the surface of water pipes – shown in Fig. 3 – covering multiple stages from mild to severe corrosion. Corrosion data is often characterized by irregular spots and depressions, and images are characterized by high contrast and noise. Corrosion defects also behave differently on different pipe materials, and the following is a partial sample of this dataset, as shown in Table II.

Fig. 3. Water pipe corrosion.

TABLE I. WATER PIPE CRACKS DATASET

Number	Image file name	Image resolution	Crack type	Crack width (MM)	Crack length (MM)
1	polie1379.png	1024x768	linear crack	0.5	15
2	polie517.png	1024x768	branch crack	0.7	12
3	polie536.png	1280x960	linear crack	0.6	20
4	polie1490.png	1280x960	branch crack	1.0	30
5	polie1542.png	1024x768	linear crack	0.4	10

TABLE II. WATER PIPE CORROSION DATASET

Number	Image name	Image resolution	Corrosion degree	Corrosion area (%)	Surface roughness (RA, MM)
1	fushi22.png	1024x768	severe	85%	2.0
2	fushi984.png	1280x960	severe	90%	2.0
3	fushi18.png	1024x768	medium	20%	2.0
4	fushi517.png	1280x960	medium	40%	0.7
5	fushi45.pngg	1024x768	severe	20%	2.0

Through the analysis performed earlier on two types of datasets, it can be found that the characteristic manifestations of various types of water pipe defects are significantly different, and the rational use of these datasets will provide a solid foundation for subsequent model development and defect detection. At the same time, the defect sample distribution and annotation details in the dataset also provide rich research materials for the design of optimized detection algorithms.

C. Tables to Support Experimental Design

1) Dataset segmentation table: The dataset segmentation table is used to display the segmentation strategy for the dataset of self-managed water pipe defects. With a reasonable proportion of training, validation, and test sets, the model can be trained and validated on different data. The specific segmentation scenarios are shown in the following Table III.

TABLE III. SEGMENTATION OF THE DATASET

Dataset name	Data volume (table)	Percentage (%)
Training Set	2202	60%
Validation Set	734	20%
Test Set	734	20%
Total	3670	100%

Table III presents the number and percentage of the training set – that is used for the training of the model, the number and percentage of the validation set – that is used for hyper-parameter tuning, and the number and percentage of the test set – that is used for the evaluation of the final model performance. This segmentation ensures that the data is fully utilized and reduces the risk of overfitting.

2) Algorithm performance indicator record table: The algorithm performance indicator record table is used to evaluate the performance of the different models used, especially in the water pipe defect detection task. Table IV shows the performance of each model in the task of detecting defects in water pipes through specific performance indicators, which is convenient for subsequent comparison and selection of the best model.

TABLE IV. ALGORITHM PERFORMANCE INDICATOR RECORD

Model name	Accuracy	Recall	F1 point	Average detection time（ms）
YOLOv5	0.90	0.85	0.87	30
Faster R-CNN	0.88	0.80	0.84	120
CNN	0.86	0.78	0.82	50

3) Model performance evaluation: The Model Performance Evaluation Form is used to provide a comprehensive evaluation of the final selected model, including how the model performs on the test set. The specific assessment results are shown in Table V.

Table V provides systematic support for the design and implementation of this experiment, covering many aspects such as the segmentation of the data set, the recording of experimental variables, the evaluation of algorithm performance, and the summary of the final results, which not only provides intuitive data support for the experimental results, but also provides a basis for subsequent model optimization and improvement. This enables an efficient and reliable water pipe defect detection system.

TABLE V. PERFORMANCE ANALYSIS

Model name	Test set sample size	Correct detection number	Number of error detections	Accuracy (%)	Recall (%)
CNN	200	170	30	85.0	80.0
YOLOv5	200	180	20	90.0	85.0
Faster R-CNN	200	160	40	80.0	75.0

D. Experimental Process

1) CNN model training details: Before the model is trained, the data is first fully prepared and preprocessed to ensure that the data provides useful features in the training and enables the model to learn effectively. The dataset is divided into 60%-20%-20% data, with 60% of the data used for training, 20% for validation, and the remaining 20% for testing. In terms of data preprocessing, a uniform resize was first applied to all images, and the dimensions of all images were adjusted to 640×640 pixels. In order to improve the robustness of the model, the images in the training set were augmented with data, including rotation, cropping, scaling, and brightness change operations, which could effectively enhance the model's ability to detect water pipe defects in different environments. Through these preprocessing and enhancements, the dataset has been significantly improved in terms of quantity and quality.

2) YOLOv5 training details with hyper-parameter tuning: In this study, the task of water pipe defect detection requires the use of efficient and accurate deep learning models. YOLOv5 was selected as the main detection model, which has high real-time performance and good detection accuracy. YOLOv5 is a single-stage detection model based on Convolutional Neural Network (CNN), which achieves a good balance between processing speed and accuracy compared to traditional multi-stage detection models. The v5.0 version of YOLO was used as the base model, and its architecture was partially customized according to the needs of the actual task. The YOLOv5 model consists of four main components:

a) Input: This is to handle the input image. Here, the image is converted to a fixed size image by resizing. Various data augmentation techniques are also applied to improve the robustness and performance of the model.

b) Backbone: This is used for feature extraction, CSPDarknet53 is used as the backbone network, which can effectively extract high-level image features.

c) Neck: This is for feature fusion, PANet (Path Aggregation Network) is used to optimize the effect of feature fusion and enhance the ability of small object detection.

d) Head: This is used to detect outputs, combined with the design of YOLO heads, it can output classification results, bounding box regression values, and confidence levels, as shown in Fig. 4.

Fig. 4. Sample of classification result.

3) Faster R-CNN of the training process: The choice of hyper-parameters is critical to the effectiveness of the model's training. The following hyper-parameter settings were used during model training.

4) Learning rate: During the training process, the setting of the learning rate has an important impact on the convergence speed and accuracy of the model. The initial learning rate of 0.001 was selected, and Cosine Annealing Learning Rate Scheduling was adopted, which could dynamically adjust the learning rate during the training process to avoid the phenomenon of gradient explosion or gradient disappearance.

5) Batch size: In order to improve the training efficiency and stability, 16 was selected as the number of images per batch. This batch size enables efficient training without exceeding the GPU memory.

6) Optimizer: The AdamW optimizer is used, which combines the advantages of Adam optimizer and L2 regularization, which can effectively avoid the overfitting problem and accelerate the convergence process.

7) Weight attenuation: In order to prevent overfitting, the weight attenuation is set to 0.0005, which can effectively constrain the complexity of the model.

8) Image enhancement strategy: In addition to the basic geometric transformations (rotation, cropping), the image is also changed in color (brightness, contrast, saturation adjustment), and noise is randomly added to further improve the robustness of the model.

9) Fusion model: The water pipe defect detection system based on the fusion strategy proposed in this study combines the advantages of Convolutional Neural Network (CNN), YOLOv5 and Faster R-CNN to achieve efficient identification, precise location and real-time detection of water pipe defects. The overall process is presented in Figure 5.

As shown in Fig. 5, the overall model consists of the following key modules:

1) Input image (high-definition water pipe image): The system uses an image of the inside of the water pipe collected by a high-definition camera device or inspection robot as input. These images retain detailed information about minor defects such as cracks and corrosion, and are the basis for subsequent processing.

2) CNN: The input image of the feature extraction module is first processed by the CNN to extract multi-level image features, including edges, textures, and semantic information. CNN serves as a shared feature extraction backbone network, providing basic feature representation for subsequent detection and classification modules.

3) YOLOv5: Image features extracted by real-time object detection are fed into the YOLOv5 model. YOLOv5 is a single-stage inspection model with high detection accuracy (90.0%) and high frame rate (50 FPS), which is suitable for large-scale, real-time defect detection scenarios. It is capable of quickly outputting defect type, location, and confidence level.

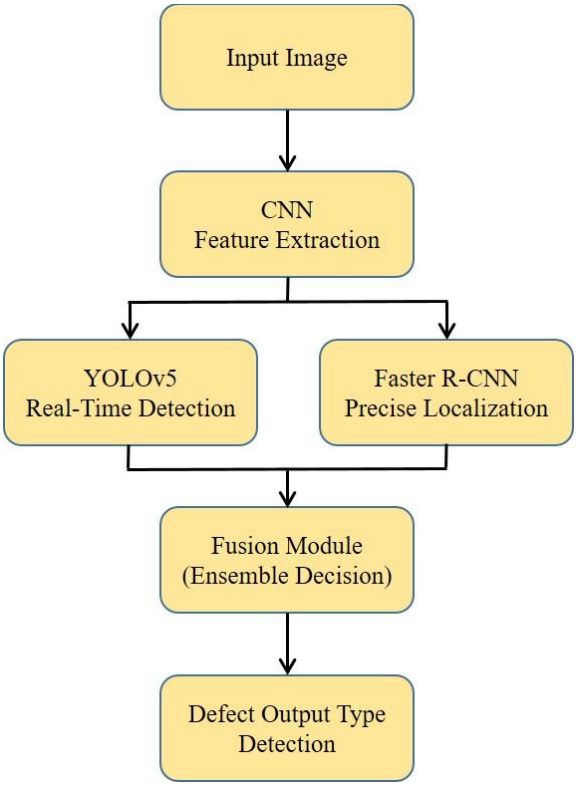


Fig. 5. Structure diagram of the fusion model.

4) Faster R-CNN: At the same time, image features are also input into the Faster R-CNN model. The model effectively improves the detection accuracy through the Regional Recommendation Network (RPN) and ROI pooling mechanism, especially for the identification of complex backgrounds, small targets or overlapping defects. It has a recall rate of 86.8% and is highly robust.

5) Fusion module: The output of YOLOv5 and Faster R-CNN will be integrated in the fusion module. By integrating factors such as scoring, position overlap and confidence, the module intelligently filters and optimizes to retain the optimal prediction results and effectively improve the overall stability and accuracy of the system.

6) Final output (defect type + location coordinates): The final output of the system includes the identified defect categories (e.g., cracks, corrosion) and their precise location

coordinates in the image. These results can be further used for network operation and maintenance decision-making, defect level assessment, and automatic maintenance path planning.

7) Ablation Study: To evaluate the contribution of each model component in the fusion system, we performed an ablation study to selectively remove each sub-model and observe its impact on detection performance shown in Table VI.

The values in Table VI confirm that YOLOv5 is a great help for real-time inspection, while Faster R-CNN improves the positioning accuracy, especially for small or overlapping defects. CNN ensure consistent feature extraction across environments. The ensemble model provides the best balance between accuracy, speed, and recall.

TABLE VI. PERFORMANCE ANALYSIS OF ABLATION EXPERIMENTS

Configuration	Accuracy	Recall	F1 Score	mAP	FPS
CNN + YOLOv5 + Faster R-CNN	92.1	85	81	85.1	45
YOLOv5	90.0	82.0	78.0	82.3	50
Faster R-CNN	89.3	86.8	72.0	81.8	32
CNN	85.0	80.0	74.0	79.5	20
YOLOv5 + CNN	91.0	84.0	80.0	84.0	42
Faster R-CNN + CNN	90.0	84.5	78.5	83.6	29

The design of this ablation test system illustrates the complementary advantages of the constituent models.

III. RESULTS

The experimental results show that there are significant differences in the performance of different models and their combinations. As shown in Table VI and Fig. 6, the fusion model (CNN + YOLOv5 + Faster R-CNN) has the best comprehensive performance, with an accuracy of 92.1%, a recall rate of 85.0%, and an F1 value of 81.0%., the mean average precision (mAP) was 85.1, and the detection speed was 45 FPS, which verified the complementary advantages of each sub-model after fusion.

Specifically, the YOLOv5 standalone model excels in real-time inspection capabilities, with an accuracy rate of 90.0% and an inspection speed of up to 50 FPS, making it ideal for real-time pipeline monitoring tasks. However, despite its slower speed (32 FPS), Faster R-CNN is more robust in complex contexts, especially in reducing false positives, missed detections, and handling target overlap, with a recall rate of 86.8%.

The ablation experiments further showed that:

1) The CNN model alone can provide basic feature extraction capabilities, but it has shortcomings in detection accuracy.

2) YOLOv5 excels at quickly and accurately identifying defects with clear and well-defined boundaries;

3) Faster R-CNN significantly improves positioning accuracy, especially when the image background is complex or there is a lot of interference.

4) It is worth noting that the fusion model not only retains the speed advantage of YOLOv5, but also introduces the localization ability of Faster R-CNN and the feature stability of CNN, which shows better performance in both standard and complex scenarios, especially in small defect detection and high-noise environments.

In summary, while YOLOv5 strikes a good balance between speed and accuracy, the fusion strategy is still better than any single model configuration across the board. Therefore, the detection system is particularly suitable for the actual pipe network inspection task, and has a broad application prospect in the application scenarios that require the coexistence of high real-time performance and high robustness.

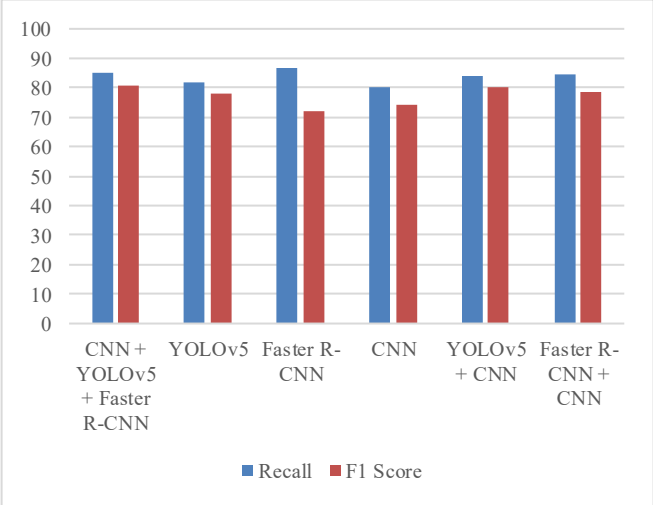


Fig. 6. Performance analysis of the fusion model.

IV. DISCUSSION

This study systematically evaluated and compared three deep learning models for water pipe defect recognition, demonstrating their complementary strengths. CNN excelled at multi-level feature extraction and, after parameter tuning, achieved 98.9 % classification accuracy on the validation set, underscoring its fundamental role in image feature analysis. As a lightweight, single-stage detector, YOLOv5 struck an effective balance between accuracy and speed, delivering 92.1 % accuracy at 45 fps – making it ideal for scenarios requiring real-time feedback. In contrast, Faster R-CNN leveraged its Region Proposal Network (RPN) and precise bounding-box regression to achieve an 86.8 % recall rate for small or overlapping defects in noisy, complex environments, highlighting its robustness against occlusion and background interference. Ablation experiments confirmed that fusing all three models outperformed any single or dual-model combination, yielding a 92.1 % accuracy, 85.0 % recall, an F1 score of 81.0, and an mAP of 85.1 %, all while maintaining 45 fps. This validates that multi-model integration can effectively combine individual advantages to boost detection precision and stability.

Despite the strong performance on binary defect detection tasks, there remains room for improvement. Future work could adopt a multi-task learning framework to jointly model defect detection, type classification, and severity assessment.

Incorporating multimodal data—such as infrared, acoustic, and thermal imaging—may further enhance adaptability in complex operating conditions. To support large-scale deployment, the system should integrate real-time feedback, dynamic optimization, and incremental learning capabilities. Moreover, embedding reinforcement learning mechanisms could enable autonomous pipeline inspection path planning and intelligent defect-repair suggestions, advancing toward a fully autonomous, intelligent, and closed-loop maintenance solution.

V. Conclusion

The proposed CNN + YOLOv5 + Faster R-CNN fusion framework successfully balances detection accuracy and real-time performance in water pipe defect recognition. Experimental results show that the integrated model surpasses individual models in key metrics – accuracy, recall, and F1 score – while sustaining a 45 fps detection speed. These findings confirm the potential of deep learning ensemble methods for intelligent, large-scale pipeline inspection and lay a solid foundation for future smart infrastructure management in urban environments.

References

[1] S. N. Betgeri, S. R. Vadyala, J. C. Matthews, and H. Lu, "Wastewater pipe defect rating model for pipe maintenance using natural language processing," Frontiers in Water, vol. 5, 2023, doi: 10.3389/frwa.2023.1123313.

[2] C. Zhao, C. Hu, H. Shao, Z. Wang, and Y. Wang, "Towards Trustworthy Multi-Label Sewer Defect Classification via Evidential Deep Learning," in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2023. doi: 10.1109/ICASSP49357.2023.10096569.

[3] B. Ha, B. Schalter, L. White, and J. Köhler, "Automatic Defect Detection in Sewer Network Using Deep Learning Based Object Detector," 2023. doi: 10.5220/0011986300003497.

[4] L. Long, H. Yang, Y. Zhou, and Y. Yang, "Research on Seismic Connectivity Reliability Analysis of Water Distribution System Based on CUDA," Water (Switzerland), vol. 15, no. 11, 2023, doi: 10.3390/w15112087.

[5] S. Wu and J. Liu, "A Multi-Scale CNN-BiLSTM Framework with An Attention Mechanism for Interpretable Structural Damage Detection," Infrastructures (Basel), vol. 10, no. 4, Apr. 2025, doi: 10.3390/infrastructures10040082.

[6] S. W. D. Turner et al., "Comparison of potential drinking water source contamination across one hundred U.S. cities," Nat Commun, vol. 12, no. 1, 2021, doi: 10.1038/s41467-021-27509-9.

[7] D. Yang, X. Zhang, T. Zhou, T. Wang, and J. Li, "A Novel Pipeline Corrosion Monitoring Method Based on Piezoelectric Active Sensing and CNN," Sensors, vol. 23, no. 2, 2023, doi: 10.3390/s23020855.

[8] M. M. Santos, A. V. Ferreira, and J. C. G. Lanzinha, "The Possibilities of Capturing Rainwater and Reducing the Impact of Floods: A Proposal for the City of Beira, Mozambique," Sustainability (Switzerland), vol. 15, no. 3, 2023, doi: 10.3390/su15032276.

[9] Z. Tang, "Analysis of Leakage Status of Urban Water Supply Network and Research Progress of Leakage Control Technology," Advances in Environmental Protection, vol. 12, no. 02, 2022, doi: 10.12677/aep.2022.122029.

[10] L. Romero-Ben, D. Alves, J. Blesa, G. Cembrano, V. Puig, and E. Duviella, "Leak detection and localization in water distribution networks: Review and perspective," Annu Rev Control, vol. 55, 2023, doi: 10.1016/j.arcontrol.2023.03.012.

[11] I. Essamlali, H. Nhaila, and M. El Khaili, "Advances in machine learning and IoT for water quality monitoring: A comprehensive review," 2024. doi: 10.1016/j.heliyon.2024.e27920.

[12] X. Liu and J. Fan, "Research on Nondestructive Testing Technology for Drilling Risers Based on Magnetic Memory and Deep Learning," Sustainability, vol. 16, no. 17, p. 7389, Aug. 2024, doi: 10.3390/su16177389.

[13] H. Wu et al., "Leveraging Optical Communication Fiber and AI for Distributed Water Pipe Leak Detection," IEEE Communications Magazine, vol. 62, no. 2, 2024, doi: 10.1109/MCOM.003.2200643.

[14] Z. Yu, H. Peiyu, Y. Bo, Y. Zhibin, L. Dongjie, and D. Guoqi, "Design and Motion Simulation of a Soft Robot for Crawling in Pipes," Appl Bionics Biomech, vol. 2023, 2023, doi: 10.1155/2023/5334604.

[15] S. N. Zainurin et al., "Advancements in Monitoring Water Quality Based on Various Sensing Methods: A Systematic Review," 2022. doi: 10.3390/ijerph192114080.

[16] M. Trigka and E. Dritsas, "A Comprehensive Survey of Deep Learning Approaches in Image Processing," Jan. 01, 2025, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/s25020531.

[17] D. Calderon and M. Najafi, "Integrated Approach to Optimizing Selection and Placement of Water Pipeline Condition Monitoring Technologies," Apr. 02, 2025. doi: 10.20944/preprints202504.0215.v1.

[18] E. Farah and I. Shahrour, "Water Leak Detection: A Comprehensive Review of Methods, Challenges, and Future Directions," Oct. 01, 2024, Multidisciplinary Digital Publishing Institute (MDPI). doi: 10.3390/w16202975.

[19] G. Michael, E. Q. Shahra, S. Basurra, W. Wu, and W. A. Jabbar, "Real-Time Pipeline Fault Detection in Water Distribution Networks Using You Only Look Once v8," Sensors, vol. 24, no. 21, Nov. 2024, doi: 10.3390/s24216982.

[20] R. Zhang et al., "ABCDWaveNet: Advancing Robust Road Ponding Detection in Fog through Dynamic Frequency-Spatial Synergy," Apr. 2025, [Online]. Available: http://arxiv.org/abs/2504.05112

[21] Y. Su et al., "Research on Intelligent Identification Technology for Bridge Cracks," Infrastructures (Basel), vol. 10, no. 5, p. 102, Apr. 2025, doi: 10.3390/infrastructures10050102.

SD-CNN: A Novel Lightweight Convolutional Neural Network Model for Fall Detection

Han-lin Shen[1], Tian-hu Wang[2]*, Hong Mu[3]
School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou, China [1, 2]
Changzhou Xutec System Technology Co., Changzhou, China[3]

***Abstract*—Aiming at the traditional deep learning fall detection model due to high computational complexity and a large number of parameters, this study proposes a lightweight convolutional neural network model, SD-CNN (SMA-Enhanced Depthwise Convolutional Neural Network), for fall detection. The model is first designed with an SMA attention module to enhance feature representation. Then, depth separable convolution is used to significantly reduce the model complexity. Finally, batch normalisation and Dropout regularisation techniques are combined to efficiently extract spatial-temporal features from temporal signals for accurate classification of fall and non-fall behaviours. The experiments use a sliding window to extract discrete features, three-axis acceleration, and synthetic acceleration as feature inputs. SD-CNN achieves 99.11% accuracy, 98.78% specificity, and 99.39% sensitivity on the homemade dataset Act, which are improved by 7.14%, 6.42%, and 9.38%, respectively, compared to CNN, while the number of parameters is reduced significantly. The effectiveness of the model is also verified by generalisation experiments on the public datasets SisFall and WEDAFall. The SD-CNN algorithm can efficiently complete the fall detection task, and the lightweight design makes it particularly suitable for wearable devices, which provides a highly efficient and reliable solution for real-time fall detection, and it has an important value for practical applications.**

Keywords*—Fall detection; lightweight; SMA attention; depth-separable convolution

I. Introduction

In the context of global ageing, falls are frequent and serious in the elderly. Not only is it the main cause of casualties, but it also often leads to hospitalisation and long-term rehabilitation, placing a heavy strain on society and the healthcare system [1]. Therefore, how to effectively detect and prevent fall events is of great relevance in the field of smart health monitoring [2].

Existing fall detection methods fall into three main categories: Machine vision-based inspection, Environmental sensor-based detection, and wearable device-based detection [3]. Machine vision-based detection relies on videos or images to determine behaviour. For example, Nana Vo et al. [4] used a lightweight network structure based on Yolov5s and OpenPose to identify video content with an accuracy of 95.43%. Mengdi Cao et al. [5] proposed a fall detection algorithm for complex scenes based on improved Yolov8, which uses video streams to identify fall behaviour in real-time, and validated on multiple video streaming datasets, achieving an average accuracy of 75.8%. However, such methods are limited in practical application by the detection range and device privacy, and cannot be widely spread. Detection based on environmental sensors relies on external sensors to obtain information about environmental changes. Wang Zhaojun et al. [6] designed an infrared array sensor system to identify human behaviour through temperature changes and used the KNN algorithm to achieve fall detection, with an accuracy rate of about 95.2%, but this approach is prone to receive interference from external environmental transformations. Detection based on wearable devices is more common. Liu Bo et al. [7] proposed a radial basis function (RBF) neural network to achieve fall detection using accelerometer data with 98.1% accuracy. Wei Jiaxue et al. [8] used an improved temporal convolutional network (TCN) to solve the problem of traditional RNN and CNN training models are complex and prone to gradient explosion phenomenon, the accuracy, sensitivity and specificity on the public and fusion datasets reached 99.43%, 98.70% and 99.34%, respectively, but the number of model parameters is relatively large, which is not conducive to the deployment on embedded devices.

To address this problem, this study proposes a lightweight fall detection model, SD-CNN, which employs Signal Modulating Attention (SMA) for processing time-series data collected by accelerometers, and which approximates the traditional self-attention mechanism [9] approach to capture long-time dependencies in the sequences through Random Feature Mapping (RFM). The use of depth-separable convolution significantly reduces the amount of computation, reduces the complexity of the feature extraction process, reduces the model parameters, and improves the computational speed, thus effectively applying to the task of fall detection.

*Corresponding author.

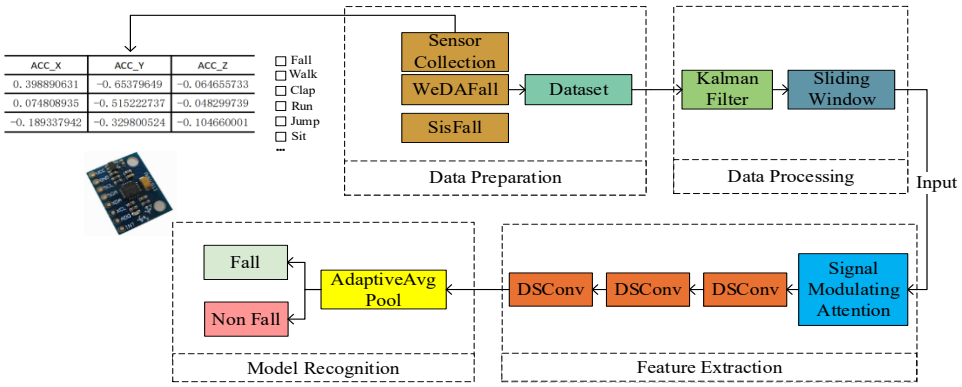


Fig. 1. General flow of the SD-CNN fall detection algorithm.

II. Related Work

A. SD-CNN Model

The key points of building a fall detection model are the appropriateness of feature extraction and the reasonableness of the constructed network model. The traditional convolutional network model is widely recognised for dealing with image problems, but due to the limitation of its convolutional kernel size, deeper networks are usually required to achieve more satisfactory results when dealing with long time series data. To address such problems, this study successively introduces the SMA attention mechanism and depth-separable convolution. SMA attention can achieve efficient and accurate fall detection on long time series data by optimising feature extraction and reducing the number of input channels. Deep separable convolution extracts feature points by decomposing the standard convolution into two independent operations, which significantly reduces the amount of computation and the number of parameters while extracting local features in each channel. The SD-CNN model consists of three depth-separable convolutional blocks, each containing a channel-by-channel convolutional layer and a point-by-point convolutional layer. The convolution kernel sizes are all set to 3, with a step size of 1 and padding of 1 to keep the time series length consistent. The number of output channels for the three convolution blocks is 32, 32, and 64 in that order. Each convolutional block is followed by batch normalisation, ReLU activation function, and average pooling (kernel size 2, step size 2) in order to progressively compress the feature length and improve feature stability. In terms of input feature design, this study divides the three-axis acceleration sensor data into fixed-length segments in a sliding-window fashion, with each sample containing 120 time steps, three channels (x, y, and z direction acceleration), and the final input dimensions of the model are 3×128. The fall detection task mainly consists of four parts, namely, data acquisition, data preprocessing, feature extraction, and model inference, and the overall layout is shown in Fig. 1.

B. SMA Attention

Although the traditional dot product attention mechanism can highlight important features of the input data and suppress unimportant features to improve the efficiency and accuracy of feature extraction, it has a computational complexity of $O(n^2)$, and the memory requirement grows by a square as the length of the sequence increases, which is particularly unfavourable for deployment in embedded devices. Therefore, in this study, we design the SMA attention mechanism, which combines the RFM idea of the EA attention mechanism [10] to reduce the computational complexity to O(n). RFM essentially provides an efficient form of approximate integration to capture long-range dependencies, where the contribution of each time step is accumulated in the global attention score. The A_{raw} obtained after normalisation characterises the distribution of importance of each time step feature in the global context, i.e, Eq. (1):

$$A_{raw} = \sum_{l=1}^{L} Q_{rf}[:, l, :] \odot K_{rf}[:, l, :] \quad (1)$$

where, the contribution of each time step is accumulated in the global attention score. A_{raw} obtained after normalisation characterises the distribution of the importance of each time step feature in the global context; larger A_{raw} correspond to signals with a consistently high response over multiple time steps (indicating that the feature is stable and important over time in the time series); smaller A_{raw} correspond to noise or short-term fluctuations (automatically suppressed by the model). The structure of the SMA attention mechanism is shown in Fig. 2.

For input feature mapping $X \in \mathbb{R}^{b\times C\times L}$, Channel dimensionality reduction is first performed by two 1×1 convolutions [Eq. (2) and Eq. (3)]:

$$Q = \mathrm{Conv1d_q}\mathbb{R}^{b\times C_r\times L} \quad Q \in \mathbb{R}^{b\times C_r\times L} \quad (2)$$

$$K = \mathrm{Conv1d_k}\mathbb{R}^{b\times C_r\times L} \quad K \in \mathbb{R}^{b\times C_r\times L} \quad (3)$$

where, b is the batch size, C is the number of input channels, L is the length of the sequence, Cr is the number of channels after dimensionality reduction, and r is the ratio of reduction.

Then, Q^T and K^T are projected into a random feature space using a random feature matrix $R_F \in \mathbb{R}^{C_r\times F}$, and ReLU activation is applied, resulting in Eq. (4) and Eq. (5):

$$Q_{rf} = \mathrm{ReLU}(Q^T \odot R_F) \in \mathbb{R}^{b\times L\times F} \quad (4)$$

$$K_{rf} = \mathrm{ReLU}(K^T \odot R_F) \in \mathbb{R}^{b\times L\times F} \quad (5)$$

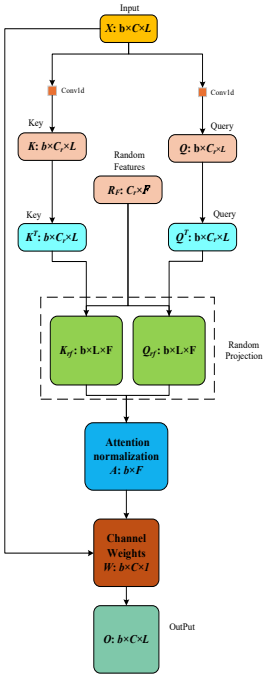


Fig. 2. Structure of the SMA attention mechanism.

where, Q^T and K^T denote the transposition of the channel dimension and the sequence length dimension, F is the number of random features. In the random feature space, Attention scores A_{raw} and normalisation factor N are then calculated as follows [Eq. (6) and Eq. (7)]:

$$A_{raw} = \sum_{l=1}^{L} Q_{rf}[:, l, :] \odot K_{rf}[:, l, :] \in \mathbb{R}^{b \times F} \quad (6)$$

$$N = \sum_{l=1}^{L} K_{rf}[:, l, :] \odot K_{rf}[:, l, :] \in \mathbb{R}^{b \times F} \quad (7)$$

where, $\odot$ is element-by-element multiplication. Then find the normalised attention fraction A, as in Eq. (8):

$$A = \frac{A_{raw}}{N+\varepsilon} \in \mathbb{R}^{b \times F} \quad (8)$$

where, ε is a small constant for numerical stability. The attention scores are further extended to all channels and averaged over the feature dimensions [Eq. (9)]:

$$W = \frac{1}{F} \sum_{f=1}^{F} A[:, f] \in \mathbb{R}^{b \times C \times 1} \quad (9)$$

Finally, the calculated weights are applied directly to the original input, as in Eq. (10):

$$O = X \odot W \in \mathbb{R}^{b \times C \times L} \quad (10)$$

C. Depthwise Separable Convolution

Depthwise Separable Convolution (DSC) consists of depth-by-depth convolution and point-by-point convolution, where depth-by-depth convolution is used to extract spatial features and point-by-point convolution is used to extract channel features [11]. Depth-separable convolution groups convolutions in feature dimensions, perform independent depth-by-depth convolution for each channel, and aggregate all channels using a 1x1 convolution (pointwise convolution) before output. This approach reduces the convolutional dimension and can effectively reduce the convolutional kernel parameters and unnecessary computations. The DSC convolutional network structure is shown in Fig. 3.

The input feature map size in Fig. 3 is DE×DF, the number of input channels is M, the size of the convolution kernel is DK, and the number of output channels is N. If the same number of channels is output using an ordinary convolutional product network, the number of parameters is DK×DK×M×N, whereas when using a depth-separable convolutional network, the number of parameters is DK×DK×M+M×N, The number of parameters is only $\frac{1}{N} + \frac{1}{D_k^2}$ times that of a normal convolution. As DK and N increase, the number of covariates of DSC relative to conventional CNN decreases dramatically.

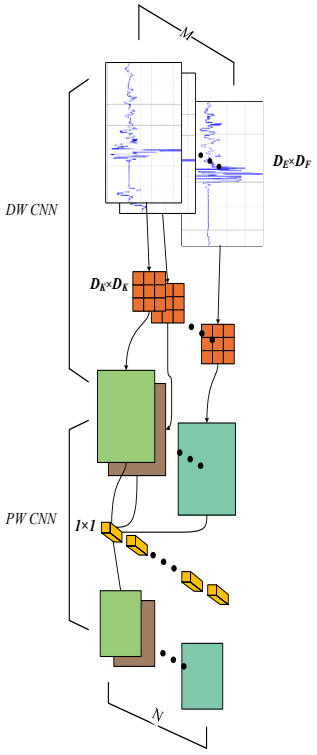


Fig. 3. Depth-separable convolutional network structure.

III. Experimental Data and Analysis

A. Datasets and Experimental Environments

There are three datasets in this study. The first one is the homemade dataset Act, and the data in the Act dataset is collected by the mpu6050 sensor. The stm32f103c8t6 is used as the main control chip, and the data sampling frequency is 50Hz. The collection was completed by 12 volunteers, who consisted of 12 young people on whom the sensor was mounted on the left wrist. The dataset contains 5 fall activities (falling backwards, falling forwards and landing on the knee, falling forwards, failing to get up from a sedentary position resulting in a fall, and falling from a standing position) and nine non-falling activities (walking, jogging, sitting down, trotting, going up the stairs, going down the stairs, opening and closing doors, applauding, and jumping), a total of 4,200 data

items are included. The dataset training set, test set and validation set are divided in the ratio of 8:1:1.

The remaining two datasets are the public datasets WEDAFall and SisFall [12]. The WEDAFall dataset equipment samples primarily at 50Hz, and data collection was done by 31 volunteers, the volunteers consisting of 20 young people and 11 older people, collected data via sensors placed on their wrists and contained a total of 3,719 pieces of data on 8 fall activities and 11 non-fall activities. The SisFall dataset device had a sampling frequency of 200Hz, and data collection was done by 38 volunteers, consisting of 23 young people and 15 older people, who collected data via sensors placed on their waists, containing a total of 4,510 pieces of data, capturing 15 fall activities and 19 non-fall activities. The specific data distribution is shown in Table I.

The deep learning framework for the experiments in this study is Pytorch, the training environment is a 64-bit Windows 11 operating system, the CPU is i7-12650H, the graphics card uses RTX-4060, the RAM is 16GB, and the code running environment is Python 3.8.

TABLE I. DATASET INFORMATION

	Non-Fall	Fall
Act	2000	2200
WEDAFall	2319	1400
Sisfall	2935	1575

B. Performance Indicators

For fall detection, the accurate identification of fall events is very important, and there are three evaluation metrics used in this study, which are accuracy, sensitivity, and specificity. Accuracy is an important measure of the model's overall classification ability, indicating the number of samples correctly predicted by the model as a proportion of the total number of samples. The equation is as follows [Eq. (11)]:

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

where, TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative.

Sensitivity and specificity, on the other hand, respond to the model's ability to distinguish between positive and negative samples. The equation is as follows [Eq. (12) and Eq. (13)]:

$$SP = \frac{TN}{TP+FN} \tag{12}$$

$$SE = \frac{TP}{TN+FP} \tag{13}$$

where, SP is specificity and SE is sensitivity, specificity and sensitivity are complementary to each other, the higher the sensitivity indicates that the model has a higher recognition rate for fall events, and the two together measure the model's ability to discriminate on positive and negative class samples.

C. Data Preprocessing

The raw triaxial acceleration components (*ax, ay, az*) are projections to the sensor's own coordinate axes and are highly dependent on its spatial attitude. Using this data directly, the model has to learn not only the motion itself, but also the complex relationship between the motion patterns and the changing sensor poses. This double learning burden increases model complexity and may impair generalisation capabilities, especially in conditions where sensor poses are unknown or changing. According to vector theory, the modulus of a vector does not change with the rotation of the coordinate system [13]. In other words, no matter how the sensor is rotated, the calculated acceleration magnitude S is constant as long as the magnitude of the total acceleration it senses remains constant.

Therefore, this study uses the acceleration amplitude as the feature, and this approach can effectively eliminate the interference caused by the sensor attitude change, making the feature robust to the sensor installation position and orientation. Acceleration amplitude indicates the size of the acceleration. The amplitude of the acceleration of the human body movement state reflects the intensity of the human body movement, which is defined as shown in Eq. (14):

$$S = \sqrt{a_x^2 + a_y^2 + a_z^2} \tag{14}$$

1) Sliding window algorithm: To extract the temporal characteristics of the data, a sliding window strategy is used. For dynamic behaviours with long duration, a larger window allows for complete capture of their signals [14]. However, blindly increasing the window size is not optimal and may lead to increased computational complexity or failure to accurately capture behavioural changes at shorter times. Therefore, a reasonable choice of window size is essential to balance the completeness of timing information with computational efficiency.

A complete fall consists of three main phases: the imbalance phase, the falling phase, and the contact with the ground phase, as shown in Fig. 4. The duration of these phases and is approximately 2s to 3s, respectively, in order to unify the size of the input data, this study makes the window able to cover a complete information of the fall event by setting the window size to 120 and the window sliding step to 10, and splits the time series into multiple fixed-length sub-sequences through the window sliding, and this method overcomes the situation that different behaviours may span different time periods, and at the same time, the model provides a sufficiently large number of training samples, which makes full use of the time series data characteristics and enables the model to capture long-time dependencies.

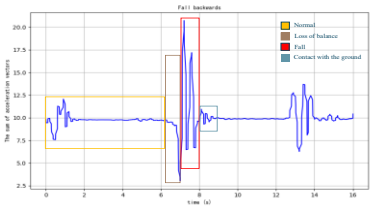


Fig. 4. Analysis of backward falling activities.

2) Low-pass filtering: Accelerometer sensors are prone to receive external noise interference when collecting data, and it

can be seen through spectral analysis that accelerometer noise is usually high-frequency noise. The frequency of human motion signals is low, with the frequency range of daily activities such as walking and running typically ranging from 0.5 Hz to 5 Hz, and the frequency of instantaneous acceleration changes for falls typically ranging from 1 Hz to 20 Hz, as shown in Fig. 5.

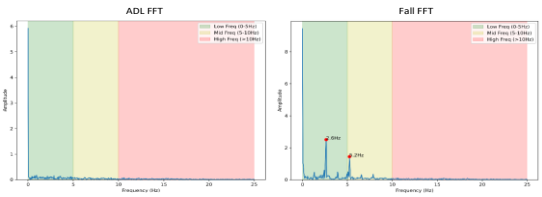


Fig. 5. Spectral analysis of human activity FFT plot.

In this study, a Butterworth low-pass filter is used to remove high-frequency noise from acceleration sensor data, which eliminates the need to rely on a physical model of signal generation and effectively reduces the interference of noise with subsequent detection tasks. After the cut-off frequency is processed, the data becomes smoother and retains the data features to have a better performance in training, as shown in Fig. 6.

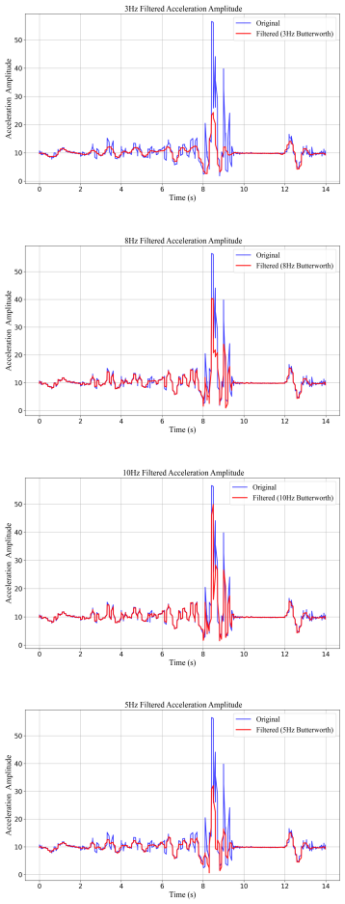


Fig. 6. Comparison of the results of Butterworth low-pass filter processing.

After comparative experiments, it was found that the model was best trained when the cut-off frequency was set to 5.5 Hz, effectively removing the interference from the sensor (see Table II).

TABLE II. COMPARISON OF TEST RESULTS

Filtering Frequency	Acc%	SP/%	SE%
\	96.34	96.25	97.11
2.5Hz	98.31	97.26	98.78
5.5Hz	**99.11**	**98.38**	**99.39**
8Hz	98.42	98.21	99.01
10Hz	97.73	96.94	98.34

IV. RESULTS

A. Comparative Experiments on Correlation Models

In order to comprehensively evaluate the performance of the proposed model and verify its effectiveness in fall detection tasks, a series of comparative experiments is designed in this study. Traditional detection methods and several state-of-the-art deep learning models were selected, including Threshold Analysis (Threshold) [15], Artificial Neural Network (ANN) [16], Convolutional Neural Network (CNN) [17], Temporal Convolutional Network (TCN) [8], Long Short-Term Memory Networks (LSTMs) [19], Gated Recurrent Neural Network (GRU) [18] and Bidirectional Long Short-Term Memory Network (Bi-LSTM) [19]. They are trained and tested on the same dataset and experimental conditions. These models were chosen because of their wide range of applications in processing time-series data and feature extraction. Through experimental comparisons, the SD-CNN model proposed in this study improves the accuracy, specificity and sensitivity by 15.31%, 16.29% and 9.38%, respectively, compared with the traditional detection method, Threshold. Compared with several other state-of-the-art deep learning models, the model in this study has the smallest number of parameters while maintaining the highest accuracy, specificity and sensitivity. This is a good indication that the SD-CNN model performs more superior in terms of classification accuracy and robustness. The specific data are shown in Table III.

TABLE III. COMPARISON OF MODELS

	Acc%	SP/%	SE%	Params	FLOPs
Threshold	83.80	89.00	83.10	/	
CNN	91.97	92.36	90.01	28480	14.6M
LSTM	93.94	92.44	95.73	46754	37.8M
GRU	95.70	87.50	96.80	47856	38.2M
Bi-LSTM	96.73	99.33	95.26	85866	42.5M
TCN	98.43	99.13	98.27	197,120	39.7M
ANN	94.02	93.20	94.21	271330	124.4M
SD-CNN	**99.11**	**98.78**	**99.39**	4076	0.97M

B. Cross-Dataset Evaluation

In order to comprehensively assess the effectiveness and usefulness of the proposed model in this study, avoid overfitting of the model to specific training data, and validate its generalisation ability on different data sources, generalisation experiments across datasets are designed and implemented in this section. In this study, two public datasets, SisFall and WEDAFall, which are widely used in the field of fall detection, are selected for testing, aiming to examine the performance of the models when confronted with new data that is different from the source of the training data. The experimental environment is consistent with "3.1" of this study. As can be seen from Table IV, SD-CNN improves accuracy, specificity, and sensitivity by 6.92%, 6.23% and 7.34%, respectively, over CNN on the WEDAFall dataset. On the SisFall dataset, the accuracy, specificity, and sensitivity are improved by 6.38%, 8.89% and 8.42%, respectively, over CNN.

TABLE IV. GENERALISATION EXPERIMENTS

Models	WEDAFall			SisFall		
	Acc%	*SP/%*	*SE%*	*Acc%*	*SP/%*	*SE%*
CNN	91.34	91.98	90.44	92.78	89.65	90.18
SD-CNN	**98.26**	**98.21**	**97.78**	**99.16**	**98.54**	**98.60**

V. DISCUSSION

The SD-CNN neural network architecture proposed in this study aims to achieve efficient and accurate human activity recognition, especially fall detection. The model achieves significant lightweighting at the model structure level by integrating deeply separable convolutions to reduce the computational effort and incorporating the SMA attention mechanism to dynamically focus on key features. The experimental evaluation results show that SD-CNN not only has significant advantages in terms of the number of model parameters and the expected computational resource consumption, but also outperforms other models in terms of key performance indicators, including 99.11% accuracy, 98.78% specificity, and 99.39% sensitivity. This proves the effectiveness of the lightweight architecture proposed in this study. Nevertheless, this study still has some limitations. Although SD-CNN shows good computational efficiency on resource-constrained devices, it still needs to be further optimised in scenarios with extreme low power consumption or higher real-time requirements. Second, the design of the attention mechanism in this study is still a relatively simplified and approximate form, and its performance differences in multi-channel and multi-modal data have not yet been fully explored.

VI. CONCLUSION

The main contribution of this research is to provide a lightweight activity recognition model, SD-CNN that balances efficiency, performance, and robustness, which is capable of recognising human activities effectively, especially in fall detection tasks. In the future, we will continue to explore this model by extending it to multimodal sensor data fusion, investigating more efficient attentional mechanisms, and validating it in larger and more diverse groups of people and scenarios to further advance its translation into practical applications.

ACKNOWLEDGMENT

This work is one of the phased achievements of Jiangsu Province Postgraduate Practical Innovation Project (SJCX24_1796), Changzhou Science and Technology Project (CZ20230025, CZ20250010, CJ20241070), and Jiangsu University of Science and Technology Postgraduate Practical Innovation Plan Project (XSJCX_46).

REFERENCES

[1] K. Durga Bhavani and M. Ferni Ukrit, "Design of inception with deep convolutional neural network based fall detection and classification model," Multimedia Tools and Applications, vol. 83, pp. 23799–23817, 2024.

[2] S. K. Gharghan and H. A. Hashim, "A comprehensive review of elderly fall detection using wireless communication and artificial intelligence techniques," Measurement, vol. 226, Art. no. 114186, 2024.

[3] Y. Li, P. Liu, Y. Fang, et al., "A Decade of Progress in Wearable Sensors for Fall Detection (2015–2024): A Network-Based Visualization Review," Sensors, vol. 25, p. 2205, 2025.

[4] F. Nana, L. Daming, C. Xiaoting, et al., "A fall detection algorithm based on the lightweight OpenPose model," Sensors and Microsystems, vol. 40, pp. 131–134, 2021.

[5] C. Mengdi, Y. Mengfan, W. Liuyi, et al., "A fall detection algorithm for complex scenes based on improved YOLOv8," Modern Information Technology, vol. 9, pp. 66–71, 2025.

[6] Z. Wang, Z. Xu, and L. Chen, "Research on Human Behaviour Recognition System Based on Infrared Array Sensor," Infrared Technology, vol. 42, pp. 231–237, 2020.

[7] B. Liu, W. Kong, J. Xiao, and M. Wang, "Fall detection algorithm based on RBF neural network," Computer Technology and Development, vol. 32, pp. 167–172, 2022.

[8] J. Wang, G. Gao, and G. Teng, "Application of improved TCN algorithm for human fall detection," Computer Engineering and Design, vol. 44, pp. 2859–2866, 2023.

[9] C. Li, M. Liu, X. Yan, et al., "Research on CNN-BiLSTM fall detection algorithm based on improved attention mechanism," Applied Sciences, vol. 12, p. 9671, 2022.

[10] Z. Shen, M. Zhang, H. Zhao, et al., "Efficient attention: Attention with linear complexities," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis., pp. 3531–3539, 2021.

[11] R. Liqiang, H. Wang, X. Peng, et al., "Complex manoeuvre recognition based on wavelet time-frequency maps and lightweight CNN-Transformer hybrid neural networks," J. Beijing Univ. Aeronaut. Astronaut., pp. 1–24, 2025.

[12] A. Sucerquia, J. D. López, and J. F. Vargas-Bonilla, "SisFall: A fall and movement dataset," Sensors, vol. 17, no. 1, p. 198, 2017.

[13] J. Handong and J. Wengang, "Simulation Study of Attitude Solving Based on Multi-MEMS Sensor Combination," Instrumentation Technology and Sensors, vol. 8, pp. 81–85, 2020.

[14] L. Xiaoguang, J. Shaokang, W. Zihui, et al., "Real-time fall detection method based on threshold and extreme random tree," Computer Application, vol. 41, pp. 2761–2766, 2021.

[15] T. Wang, B. Wang, Y. Shen, et al., "Accelerometer-based human fall detection using sparrow search algorithm and back propagation neural network," Measurement, vol. 204, p. 112104, 2022.

[16] K. Lisa, W. Suzheng, Y.-Q. Chen, et al., "A Review of Fall Detection Algorithms Based on Wearable Devices," J. Zhejiang Univ., vol. 52, pp. 1717–1728, 2018.

[17] L. Yan, Z. Meng, J. Wuhào, et al., "Design of fall detection system for the elderly using convolutional neural network," J. Zhejiang Univ., vol. 53, pp. 1130–1138, 2019.

[18] F. Luna-Perejón, M. J. Domínguez-Morales, and A. Civit-Balcells, "Wearable fall detector using recurrent neural networks," Sensors, vol. 19, no. 22, p. 4885, 2019.

[19] M. Dong and J. Peng, "A study on wearable fall detection based on bidirectional LSTM neural network," J. Guangxi Normal Univ., vol. 40, pp. 141–150, 2022.

Efficient Lightweight Detection and Classification Method for Field-Grown Horticultural Crops

Yaru Huang[1], Hua Zhou[2], Zhongyi Shu[3]*
School of Cyberspace Security, Software Engineering Institute of Guangzhou, Guangzhou, China [1, 2]
Zhongnan University of Economics and Law, Wuhan, Hubei, China[3]

***Abstract*—As the core carrier of human food supply and agricultural economy, manual management in large-scale crop cultivation faces bottlenecks such as low efficiency, high cost, and difficulty in standardization. There is an urgent need for computer vision technology to realize automated detection and growth stage classification. However, most existing algorithms rely on high-performance GPUs for operation, resulting in high hardware costs, which makes it difficult to popularize them in low-end agricultural edge devices (e.g., embedded controllers, low-cost industrial computers). This study proposes a lightweight crop detection and classification model, Lite-CropNet. It builds a neural network architecture based on the CSPDarknet backbone network, designs a concise decoder, and adopts four-scale detection heads to adapt to crop targets of different sizes, balancing high accuracy and lightweight characteristics. Using tomatoes as the experimental object, tests on the TomatOD dataset (simulating real greenhouse environments) show that Lite-CropNet outperforms advanced methods, with a mean Average Precision (mAP)@0.5 of 85.7%. Under the conditions of the GTX 1650 GPU and 640×640 resolution, the Frame Per Second (FPS) reaches 76.9, and the model size is only 4.4M. This neural network model can efficiently complete tomato detection and maturity classification, and its architecture and design can also be transferred to crops such as potatoes and strawberries, providing a cost-effective and highly universal automated solution for agricultural production.**

Keywords*—Computer vision; neural network; object detection and classification; lightweight; horticultural crops

I. Introduction

In large-scale agricultural production, accurate detection and growth stage classification of crops are core links for yield assessment, quality control, and automated management. Whether for vegetable and fruit crops such as tomatoes and potatoes, or food crops such as wheat and corn, real-time monitoring of their growth status directly affects production efficiency and economic benefits. With population growth and the development of agricultural intensification, models such as greenhouses and large-scale field cultivation have been widely applied [1]. However, traditional manual detection methods, limited by low efficiency, long time consumption, and high cost, can hardly meet the management needs of large-scale and multi-variety crops [2], [3].

With the development of computer vision and deep learning technologies, the advancement of image recognition technology in recent years has enabled Convolutional Neural Networks (CNNs) to be widely used in deep learning [4], becoming an important method for image classification, object detection, and recognition. Especially in the agricultural field, deep learning models based on the CNN architecture have been gradually applied [5], [6], [7] due to their high accuracy in image classification and object detection tasks. At this point, how to collect these images for deep learning training is a question that researchers should consider. Currently, commonly used methods include satellite remote sensing, unmanned aerial vehicles (UAVs), fixed platforms, and unmanned vehicles. These devices have their own advantages and disadvantages: satellite remote sensing systems are suitable for long-distance and large-scale monitoring but have relatively low image resolution, making it difficult to capture objects at the small organ level [8]; UAVs are usually used for large-scale field monitoring [9]; these two technologies are suitable for long-distance shooting and have high costs. Fixed platforms are installed at fixed positions and are more suitable for monitoring small-scale areas. In contrast, unmanned vehicle technology has shown excellent performance in large-scale farmland or greenhouses—it can not only adapt to the detection of underlying fruits of hanging crops such as tomatoes but also meet the close-range image collection of low-growing crops such as potatoes, providing data support for high-precision detection of multi-variety crops.

However, existing crop detection and classification technologies still face two core challenges: First, insufficient scene adaptability—different crops have significant differences in growth forms (e.g., hanging tomatoes, creeping potatoes, clustered strawberries) and appearance features (color, shape, size). In addition, the field environment has interferences such as light changes and occlusions (overlapping of leaves, branches, and fruits), making it difficult for a single model to adapt to multi-variety crops. Second, high hardware dependence—most existing models require high-performance GPUs for support, with large parameter scales and high computational costs, which cannot be deployed in low-end agricultural edge devices (e.g., embedded controllers, low-cost industrial computers), limiting the large-scale application of multi-variety crop detection technologies. Bathini and Usha Rani [10] also pointed out in their review on the application of artificial intelligence in agricultural crop yield analysis that although current deep learning models perform well in crop management, yield prediction, and other fields, they generally have problems of high dependence on data quality and insufficient adaptability of lightweight models.

To address this challenge, this study adopts the TomatOD dataset [11], a highly specialized and innovative dataset designed specifically for tomato fruit object detection and

*Corresponding author.

classification. Collected in a real greenhouse environment, it includes low-light conditions and long-distance shooting scenarios, providing challenging scenes for the model and helping it better adapt to real-world applications. Against this background, the study proposes the Lite-CropNet model, whose core differences in mathematical architecture from existing methods are as follows: 1) Feature extraction: A multi-stage structure of $Ck(m) - Pk(m)$ is adopted to output multi-scale feature maps with 32/64/128/256 channels; 2) Loss function: Smoothed Intersection over Union (SIoU) loss (including angle cost $\Delta = \sum_{t=x,\gamma} (1 - e^{-}{}_{\gamma\varrho t})$ and shape cost $\Omega = \sum_{t=} w,h\ (1 - e^{-w}{}_{t})^{th}$) is introduced; 3) Lightweight: The parameter computation amount is controlled to 4.4M according to the formula $Params = i \cdot (k \cdot k) \cdot o + o$ (where i is input size, k is convolution kernel, o is output size), which is only 2.8% of the DETR model (158M), solving the problem of high hardware dependence. While being lightweight, Lite-CropNet maintains high accuracy, with a mean Average Precision (mAP) of 85.7%. Notably, Lite-CropNet has only 4.4M parameters, and under the image resolution of 640×640, it achieves an excellent frame rate of 76.9 Frames Per Second (FPS) on the low-cost GTX 1650 GPU. This enables Lite-CropNet to have wide applicability in real-time applications and provides a cost-effective and efficient automated solution for agricultural production.

In summary, the main contributions of this study are threefold: 1) Lite-CropNet: a novel deep convolutional network with a lightweight encoder and a decoder that effectively processes encoded features, alleviating the problems of information loss and degradation. 2) Innovative use of four detection heads of different scales, which not only has excellent detection performance for medium and large target crops such as tomatoes but also can accurately identify other small clustered crops, improving the model's crop variety adaptability. 3) High real-time efficiency is reported on low-cost devices, providing an effective and economical solution for the application of modern high-throughput plant phenotyping platforms.

The structure of this study is arranged as follows: Section I (this section) introduces the research background. Section II presents related work by researchers and highlights the problem statement. Section III provides a detailed description of the Lite-CropNet model. Section IV introduces dataset details, challenges, and experimental design. Section V conducts experiments and performs comparative analysis with other models from multiple dimensions. Section VI discusses the research. Section VII draws the conclusions of this research and proposes future work.

II. Related Work

In recent years, deep learning models based on Convolutional Neural Networks (CNNs) have been widely used in agricultural crop detection and classification tasks. Researchers have carried out explorations around "improving accuracy", "adapting to scenes", and "optimizing efficiency", forming research results in multiple directions. In terms of crop and weed detection and classification, Ahmad et al. [12] proposed a detection model based on Vision Transformer (ViT). Through pixel-level annotation training of high-resolution UAV images, it overcomes the problem of similar appearance between crops and weeds, achieving a classification accuracy of 89.4%, which is significantly better than traditional models such as UNet and Fully Convolutional Network (FCN). Zhang et al. [13] further proposed the LMS-YOLO11n lightweight multi-scale weed detection model. Through the design of FastGLU feature extraction module and Attention Hierarchical Feature Pyramid Network (AHFPN) feature fusion network, the mAP is increased by 2.5% on the CottonWeedDet3 dataset, while the model parameters and computational complexity are reduced by 37% and 26% respectively. Reedha et al. [14] used Transformer neural networks to process high-resolution UAV images, further optimizing the classification accuracy of weeds and crops. López-Correa et al. [15] designed an intelligent weed management system in tomato fields, providing technical support for precise weeding in the field.

In terms of crop organ and maturity detection, Mu et al. [16] proposed a tomato detection model based on Faster R-CNN and ResNet-101. By transferring learning from the COCO dataset, it can automatically detect intact green tomatoes (ignoring occlusion and growth stages), with an average precision of 87.83% on the test set. Su et al. [17] improved the SE-YOLOv3-MobileNetV1 network, introduced a channel attention mechanism, and could distinguish four tomato maturity levels, with an average precision of 87.7%. Rahim et al. [18] used Faster R-CNN to realize tomato flower detection and counting in greenhouses. Rahim et al. [19] further quantified the number of grapevine inflorescences and flowers, providing data for yield prediction. Suh et al. [20] realized the classification of sugar beets and potatoes through transfer learning, adapting to field scenarios. Nkemelu et al. [21] used a deep CNN to complete crop seedling classification, solving the problem of seedling stage recognition.

In terms of data collection, commonly used devices include satellite remote sensing, UAVs, fixed platforms, and unmanned vehicles [8], [9]. Remote sensing is suitable for large-scale areas but has low resolution; UAVs have high costs; fixed platforms are suitable for small-scale areas; unmanned vehicles are suitable for image collection of various crops such as tomatoes and potatoes. Tsironis et al. [11], [22] constructed the TomatOD dataset, which contains 277 images and 2,413 tomato samples, annotates three maturity stages, and covers complex scenarios such as light changes and occlusions, with good challenging properties.

Although existing studies have made progress, there are still three major bottlenecks: First, poor scene adaptability—most models rely on ideal datasets and are difficult to cope with complex field environments such as low light and high occlusion. Second, high hardware dependence—models such as DETR [23] and Faster R-CNN [24] have large parameter sizes (DETR reaches 158M), making it difficult to deploy them on agricultural edge devices. Third, weak multi-variety adaptability—existing studies mostly focus on a single crop or task, and do not fully consider the morphological differences between crops (e.g., hanging tomatoes, clustered strawberries), making it difficult to achieve universal detection.

To address the above problems, this study proposes Lite-CropNet, achieving the following breakthroughs: In terms of lightweight design, through multi-stage feature extraction and a concise decoder, the number of parameters is controlled at 4.4M (only 2.8% of DETR), adapting to low-end devices such as GTX 1650 and J4125. In terms of scene robustness, training is conducted based on the TomatOD dataset, and the SIoU loss function (including angle and shape costs) is introduced to improve adaptability to complex field environments. In terms of multi-variety detection capability, an innovative four-scale detection head is designed to support various crops such as tomatoes (medium and large-sized) and strawberries (small clustered), reducing the cost and complexity of multi-variety deployment.

III. Materials and Methods

This section provides a detailed introduction to the design of the Lite-CropNet deep learning framework for crop detection and classification.

A. Lite-CropNet Model Design

In crop detection and classification, the model is required to have higher robustness due to the influence of various natural factors and imaging from different shooting angles. Moreover, considering the deployment needs of edge devices in the context of plant science, the model is also required to be lightweight. Therefore, in the model architecture design, both aspects are taken into account: while reducing parameters, the performance of the model is maintained, and the Lite-CropNet model is designed, which consists of three parts. The Encoder (feature extraction network) is responsible for extracting image features; after feature enhancement of the feature maps, the Decoder combines, utilizes, and decodes the features from the Encoder, aggregates low-level spatial features and high-level semantic features, and improves the recognition accuracy of objects of different scales. Finally, the features are transmitted to the Detector for detection, generating object bounding boxes with coordinates, categories, and confidence levels. The overall architecture of the Lite-CropNet model is shown in Fig. 1, and the global architecture of Lite-CropNet and its optimization are introduced below.

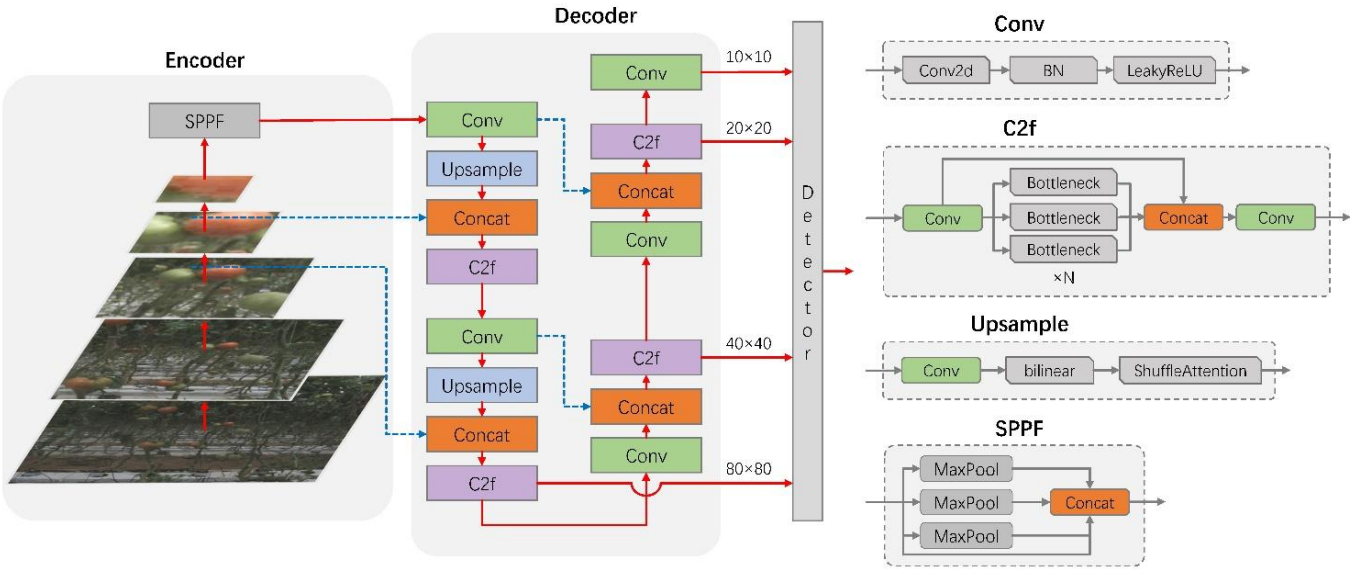


Fig. 1. Lite-CropNet model structure and module details.

1) Encoder: For the encoder part, the general structure of Cross Stage Partial Darknet (CSPDarknet) [25] is adopted and its design is retained. First, an RGB image $I \in R^{H\times W\times 3}$ is given as input, where H and W represent the height and width of the output feature map, respectively. The entire Encoder is composed of 5 convolutional layers and 4 feature extraction layers, specifically defined as Eq. (1):

$$C3(16) - C3(32) - P1(32) - C3(64) - P2(64) - C3(128) - P3(128) - C3(256) - P4(256) \quad (1)$$

Here, $Ck(m)$ denotes a 2D convolutional layer with m channels and $k \times k$ filters, all with a stride of 2; $Pk(m)$ denotes the k^{th} feature extraction layer that outputs m channels. By inserting CSPLayer [26] at different stages to extract features, 32, 64, 128 and 256 channels are output, generating feature maps at different stages. After the last feature extraction layer (P4), a Spatial Pyramid Pooling - Fast (SPPF) [27] module is introduced. The number of input channels of this module is equal to the number of output channels of the last feature layer ($C_{input} = 256$), and the kernel size is $K_{SPPF} = 5 \times 5$. Without changing the size of the feature map, pooling is performed on the feature map at different scales, which helps the network better capture the semantic information of various targets.

2) Decoder: In the design of the Decoder, first, a convolutional layer with a kernel size of 1×1 is used to compress the 256-channel feature map extracted by the SPPF in the Encoder to 128 channels, reducing computational complexity while retaining key features. Subsequently, upsampling is performed to increase the spatial resolution of

the feature map, which is then concatenated with the feature map of P4. A Cross Stage Partial Fusion (C2f) module [28] is introduced to optimize the feature representation. Next, the same operation is carried out again, and the result is concatenated with the feature map of P3. Finally, a convolution with a stride of 2 and a kernel size of 3×3 is further connected, and the result is concatenated with the previous feature maps to form a more optimized feature representation. In this process, emphasis is placed on the utilization of P2 and P3 feature maps because they have a suitable scale for object detection.

Another important design of the Lite-CropNet network is to add a prediction head to detect ultra-large targets while simplifying the process of adding detection layers. During the inspection of the dataset, it was found that some images contain ultra-large crop targets, which also exist in practical applications. Therefore, a convolution module with a stride of 2 and a kernel size of 3×3 is directly added after the third feature map passed to the head. This innovation effectively improves the detection ability of ultra-large targets without significantly affecting the overall computational complexity of the model. Through gradual feature fusion and representation optimization (without significantly affecting the overall computational complexity of the model), the decoder finally transmits four feature maps of different sizes to the detection module (Detector) to output the final object detection results. The design of the Decoder has clear objectives in each step of the operation. From channel compression, upsampling to feature concatenation, all are aimed at comprehensively utilizing the sufficiently deep encoded feature layers in the encoder to respond to more abstract information, and combining an adaptive strategy to restore spatial resolution and improve the detection accuracy of targets of different scales.

B. Activation Function

The selection of the activation function is of great significance in neural networks. It can introduce non-linear factors into the neural network and improve the model's expressive ability. In the Lite-CropNet model, attempts were made to use SiLU [29], ReLU [30], and LeakyReLU [31] activation functions, and their performances were compared. Finally, the more optimal LeakyReLU activation function was selected.

SiLU is a smooth and differentiable activation function with non-monotonicity, which can more easily capture complex patterns. Its non-linear definition is Eq. (2):

$$SiLU(x) = x \cdot \sigma(x) \tag{2}$$

In contrast, ReLU is widely used in neural networks and has simple computation, which helps to improve the computation speed of the network. It can effectively alleviate the problem of gradient disappearance and introduce sparse activation during the training process, making the network easier to optimize. Its non-linear definition is Eq. (3):

$$ReLU(x) = max(0, x) \tag{3}$$

where, x is the input. ReLU can effectively alleviate the problem of gradient disappearance and introduce sparse activation during the training process, making the network easier to optimize. However, it should be noted that ReLU has the problem of neuron death, that is, the output is zero when the input is negative, resulting in some neurons not being activated. To solve the dead neuron problem of ReLU, LeakyReLU introduces a small negative slope for negative inputs while maintaining simplicity. LeakyReLU retains the advantages of ReLU and provides a stronger gradient signal by allowing small negative gradients. This helps to avoid the problem of gradient disappearance during training while maintaining the computational efficiency of the model. The equation of LeakyReLU is Eq. (4):

$$LeakyReLU(x) = \begin{cases} x\ (x > 0) \\ \alpha x\ (x \leq 0) \end{cases} \tag{4}$$

where, α is a small positive slope, such as $0.1, 0.01$, or even smaller.

In practical tasks, the selection of the activation function usually depends on specific needs and experimental results. In the Lite-CropNet model, considering the need to enhance the learning ability, LeakyReLU was finally selected as the activation function because it solves the dead neuron problem, has a lower computational cost compared to SiLU, and has a more stable training process compared to ReLU.

C. Attention Mechanism

The attention mechanism is widely used in computer vision technology. It is a cognitive process that selectively acts on relevant information, enabling the model to ignore redundant information and focus more on useful information. Hu et al. [32] proposed the Squeeze-and-Excitation (SE) attention mechanism, which focuses on channel information and can enhance the channel features of the input feature map. However, due to the introduction of the FC fully connected layer, its computational complexity is relatively high. Woo et al. [33] proposed the Convolutional Block Attention Module (CBAM) attention module. Compared with the SE module, CBAM adds a spatial attention mechanism and focuses on regions of interest. The Efficient Channel Attention (ECA) attention mechanism [34] with global context self-attention generates weights by using global context information, allowing the model to better capture context information, but its computational complexity is relatively high. Shuffle Attention (SA) [35] generates context information by rearranging input features, enhances the robustness of the model to spatial changes, and optimizes the allocation of attention weights through feature grouping and rearrangement. Its core equation is Eq. (5):

$$F_{SA} = Shuffle(Conv2(Conv1(F_{in}, \frac{C}{g}))) \tag{5}$$

where, F_{in} is the input feature map, C is the number of input channels (such as 128, 64), g is the number of groups, $Conv1$ is a 1×1 convolution (compressing the channels to $\frac{C}{g}$), $Conv2$ is a 3×3 convolution (restoring the channels to C), and $Shuffle(\cdot)$ is the feature rearrangement operation. Yu et al. [36] proposed the Mlt_ECA attention module, which is a dimension-free local cross-channel interaction strategy. It generates feature weights by performing 1D convolution

operations, and the convolution kernel size K is adaptively determined by the mapping of the channel dimension C, which is defined as Eq. (6):

$$K = \left|\frac{\log_2(c)+\beta}{\alpha}\right|_{odd} \quad (6)$$

α and β are adjustable hyperparameters, and "odd" indicates that K takes an odd number.

Considering that most of the fruit target areas in the dataset are small and there is a lot of background, which is not helpful for the detection and classification of crop fruits. Therefore, to reduce the impact of the background and pay more attention to the required feature information, during the design process of the Lite-CropNet model, attempts were made to compare the effects of different attention mechanisms and apply them to different stages of the model. A large number of experiments were conducted, and it was found that the experimental results were better when the Shuffle Attention mechanism was applied to the upsampling and downsampling modules. The performance differences of different attention mechanisms will be verified in the ablation experiments of the article.

D. Loss Function

As a metric to calculate the error between the forward propagation result of a neural network in each iteration and the ground truth, the loss function guides the weight adjustment in backpropagation. In the implementation of Lite-CropNet, various commonly used loss functions were tested. For the bounding box loss function, the Complete Intersection over Union (CIoU) loss function was initially considered, with its calculation methods shown in Eq. (7) and Eq. (8):

$$CIoU = IoU - \frac{\rho^2\left(b,b^{g^t}\right)}{c^2} - \alpha v \quad (7)$$

$$L_{CIoU} = 1 - CIoU \quad (8)$$

Intersection over Union (IoU) represents the intersection ratio between the predicted bounding box and the ground truth box; denotes the shortest diagonal length of the minimum enclosing box that contains both the predicted box and the ground truth box; $\rho^2\left(b,b^{g^t}\right)$ is the Euclidean distance between the center of the ground truth box and the center of the predicted box; α is a positive balance parameter; and v represents the consistency of the aspect ratio between the predicted box and the ground truth box. The calculation methods of α and v are shown in Eq. (9) and Eq. (10):

$$v = \frac{4}{\pi^2}\left(arctan\frac{w^{g^t}}{h^{g^t}} - arctan\frac{w}{h}\right)^2 \quad (9)$$

$$\alpha = \frac{v}{(1-IoU)+v} \quad (10)$$

In Eq. (9), h^{g^t} and w^{g^t} represent the height and width of the ground truth box, respectively; h and w represent the height and width of the predicted box, respectively. The CIoU loss function comprehensively considers the overlapping area, aspect ratio, and center distance, and can well measure the relative position between boxes. However, CIoU does not consider the direction matching between the target box and the predicted box, leading to a slow convergence rate. Therefore, this study considered an alternative approach using the Smoothed Intersection over Union (SIoU) loss function [37]. SIoU optimizes the loss calculation by introducing the vector angle between the target box and the predicted box, and it plays an important role in strawberry detection networks through the linear combination of four components: angle cost, distance cost, shape cost, and IoU cost. Its calculation methods are shown in Eq. (11) and Eq. (12):

$$L_{SIoU} = 1 - IoU + \frac{\Delta+\Omega}{2} \quad (11)$$

$$IoU = \frac{|B \cap B^{GT}|}{|B \cup B^{GT}|} \quad (12)$$

where, B and B^{GT} represent the predicted box and the ground truth box, respectively; Ω denotes the shape cost; and Δ represents the distance cost (which incorporates the angle cost and redefines the distance metric). The equations for Ω and Δ are defined as Eq. (13) and Eq. (14):

$$\Omega = \sum_{t=w,h}\left(1 - e^{-w_t}\right)^{\theta} \quad (13)$$

$$\Delta = \sum_{t=x,y}\left(1 - e^{-\gamma\rho_t}\right) \quad (14)$$

In Eq. (13), $w_w = \frac{|w-w^{g^t}|}{max\left(w,w^{g^t}\right)}$, $w_h = \frac{|h-h^{g^t}|}{max\left(h,h^{g^t}\right)}$ and θ represents the degree of attention paid to Ω.

In Eq. (14), $\rho_x = \left(\frac{\left|b_{cx}^{g^t}-b_{c_x}\right|}{c_w}\right)^2$, $\rho_y = \left(\frac{\left|b_{cy}^{g^t}-b_{c_y}\right|}{c_h}\right)^2$, and γ is defined as:

$$p = arcsin\frac{\max\left(b_{cy}^{g^t},b_{c_y}\right)-\min\left(b_{cy}^{g^t},b_{c_y}\right)}{\sqrt{\left(b_{cx}^{g^t}-b_{c_x}\right)^2+\left(b_{cy}^{g^t}-b_{c_y}\right)^2}} - \frac{\pi}{4} \quad (15)$$

$$\gamma = 1 + 2\,sin^2(p) \quad (16)$$

In Eq. (15) and Eq. (16), $b_{c_x}^{g^t}$ and $b_{c_y}^{g^t}$ are the coordinates of the center of the ground truth box; b_{c_x} and b_{c_y} are the coordinates of the center of the predicted box.

The loss function plays a key role in evaluating the performance of a detection model. Typically, at least three loss functions need to be defined: object loss, classification loss, and bounding box loss. The bounding box loss has a significant impact on the detection accuracy and convergence speed of the network model.

To improve the stability of the model in predicting the size and position of target boxes, the SIoU loss function was introduced, which incorporates the vector angle of the expected regression into the calculation of the bounding box regression loss function. The mathematical expression of the SIoU loss function is as follows:

$$L_{SIoU} = \alpha \cdot AngleCost + \beta \cdot DistanceCost + \gamma \cdot ShapeCost + \delta \cdot IoUCost \quad (17)$$

In Eq. (17), α, β, γ and δ are the weight coefficients of the loss terms; $AngleCost$ represents the angle cost; $DistanceCost$ represents the distance

cost; $ShapeCost$ represents the shape cost; and $IoUCost$ represents the IoU cost. Through the linear combination of these four components (angle cost, distance cost, shape cost, and IoU cost), the difference between the predicted box and the ground truth box is measured more comprehensively.

In summary, by introducing the vector angle between the required regression boxes, the SIoU loss function redefines the distance loss, effectively reducing the degree of freedom of regression, accelerating the convergence speed of the network, and further improving the regression accuracy. Therefore, this study adopts the SIoU loss function as the loss function for bounding box regression, which improves the detection accuracy and stability, enabling Lite-CropNet to better complete the classification task of crop fruits of different sizes. This will be further verified in the experiments below.

IV. Experimental Design

In this section, the information of the adopted dataset is first introduced, followed by the presentation of evaluation metrics and experimental details. Additionally, the proposed Lite-CropNet model is compared with advanced detectors on the TomatOD dataset, and their performance is reported. Through the visualization of inference results and the analysis of the model's defects, the characteristics of the model are discussed in depth. Finally, ablation experiments are conducted to verify the selection of the key attention mechanism in Lite-CropNet.

A. Dataset Materials

In this experiment, the TomatOD dataset [11], publicly released by Tsironis et al. (2020) [22], was adopted. The selection basis and dataset characteristics are as follows: This dataset was collected from a real soilless tomato greenhouse on Crete, Greece, aiming to simulate the actual agricultural scenario of robotic arm navigation. Its scene authenticity, task adaptability, and comprehensive challenges are highly consistent with the core goal of this study—"developing a practical lightweight crop detection model". On the one hand, the dataset covers field interferences such as light fluctuations and occlusions by branches, leaves, and fruits, which can verify the model's robustness in complex environments. On the other hand, it contains 277 images and 2,413 tomato samples, which not only annotate accurate bounding boxes but also classify three maturity stages (unripe (green) - semi-ripe (orange-red) - fully ripe (red)) according to color phenotypes.

Since this dataset already provides bounding box annotations—each tomato is labeled with a bounding rectangle and marked with corresponding labels for the three growth stages—these annotated data form the basis for model training. To ensure the accuracy of the annotations, LabelImg [38] was used to check the tomato annotations in the dataset, and at the same time, the characteristics of the dataset were analyzed in depth to help design a better model. The number of tomatoes in each image ranges from 1 to 21, and the size of the annotation box accounts for 3% to 15% of the image size. Tomatoes with a size smaller than a certain range are considered out of scope and not annotated or detected. In terms of the number of category instances, the proportion of labels of different categories in all labels shows that the number of unripe instances is the largest, while the number of semi-ripe instances is the smallest. The TomatOD dataset uses an 8:2 ratio to divide the images into a training set and a test set. The details of the number of images and classification of the dataset are shown in Table I.

TABLE I. Dataset Details

Growth Stage	Training Set Instances	Testing Set Instances	Total	Percentage(Total)
Unripe	1,295	292	1,587	65.8%
Semi-ripe	318	77	395	16.4%
Fully-ripe	332	99	431	17.8%
Total	1,945	468	2,413	100.0%

In addition, this dataset has a unique feature: it is collected in a real greenhouse tomato field. In actual tomato planting scenarios, conditions such as light may change at any time, and the dataset takes this into account, meeting the experimental needs. It includes scenarios such as occlusion by branches, occlusion by tomatoes, occlusion by leaves, and unsuitable dim light, even the superposition of multiple situations. The dataset contains multiple types of occlusion scenarios, and the occlusion degree is quantified using the mathematical equation: $= 1 - \frac{|B_{vis}|}{|B_{gt}|}$. Among them, B_{gt} is the area of the real bounding box of the fruit, B_{vis} is the area of the visible region of the fruit, and O is the occlusion degree ($O \in [0,1]$, where $O = 0$ means no occlusion and $O = 1$ means full occlusion). As shown in Fig. 2, representative images are carefully selected to show the four main challenges in the tomato detection and classification task. For clear display, the brightness of the images has been enhanced, while the actual images have weak light. It is worth noting that to more intuitively display the key parts, these images are shown after magnification. In the actual dataset, the imaging distance of the images is longer, and the tomato targets are smaller.

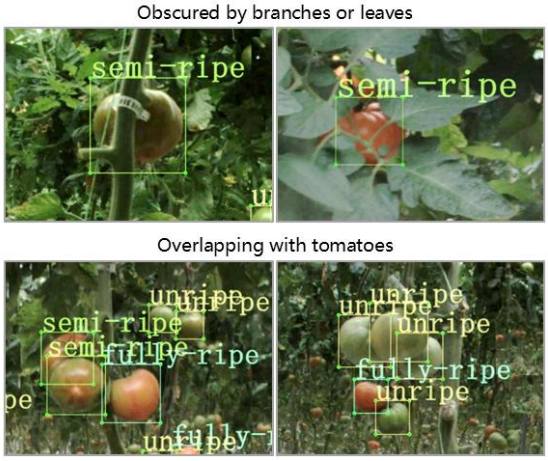


Fig. 2. Main challenges in the TomatOD dataset.

In summary, it can be seen that the many characteristics of the TomatOD dataset increase the difficulty of object detection. Through more realistic scene challenges, it helps the model

better adapt to real-world applications and achieve better robustness and generalization.

B. Experimental Configuration

This experiment was implemented using the PyTorch deep learning framework [39] and accelerated using CUDA. The TomatOD dataset was used for training and evaluation, and all 222 images in the training set were used for training and validation. The image size was adjusted to 640×640 pixels for training, which also meets the resolution requirements when deploying to low-end edge devices. During the training process, Adaptive Moment Estimation (Adam) [40] was used as the optimizer, with the initial learning rate (lr0) and cyclic learning rate (lrf) both set to 0.01, the momentum set to 0.8, and the batch size set to 8 images per batch. The hardware parameters of the equipment used in the experiment were Intel(R) Core(TM) i5-13400F and NVIDIA GeForce GTX 3090 GPU, equipped with a deep neural network acceleration library of CUDA version 11.8, parallel computing framework, and CUDNN version 8.9.5.

After the configuration of relevant parameters was completed, according to the actual convergence of the model, the training process was configured to 280 epochs for optimization. Data augmentation techniques were adopted, including random horizontal flipping, random adjustment of brightness or contrast, and random cropping. The trained model was evaluated using the test set, and each image in the test set was different from those used in the training set. Fig. 3 shows the changes in various losses and mAP values with epochs during the entire training process. It can be observed that the model was properly trained, showing convergence without overfitting.

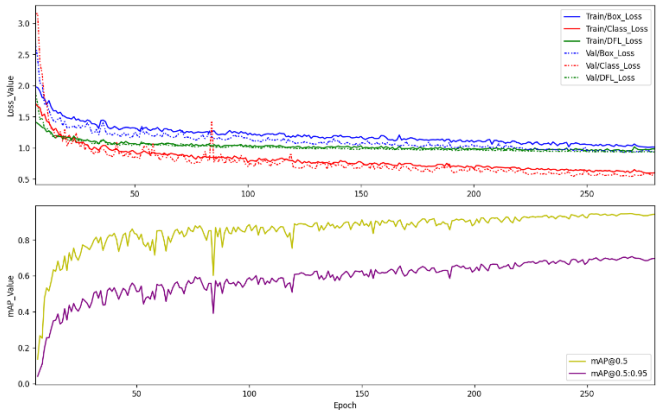


Fig. 3. Changes in loss and mAP values with epochs during training.

C. Evaluation Metrics

To evaluate the detection performance of the Lite-CropNet model, this study uses Precision (P), Recall (R), mean Average Precision (mAP@0.5, mAP@0.5:0.95), and F1 (F1-score) as evaluation metrics. A standard Intersection over Union (IoU) threshold of 0.5 was used in the experiment. If the overlap between a predicted bounding box and a labeled bounding box exceeds the IoU threshold, it is considered correct (true positive). Otherwise, the predicted bounding box is considered a false positive. When a labeled bounding box overlaps with a predicted bounding box below the threshold, it is considered a false negative. Precision (P) represents the proportion of correctly predicted objects by the model among all predicted objects. Recall (R) represents the proportion of correctly predicted objects by the model among all actual objects. Average Precision (AP) is defined as the area under the Precision-Recall (P-R) curve formed by the above Precision and Recall. The F1-score evaluates the model by balancing the weights of Precision and Recall. They are defined by the following equations, respectively [see Eq. (18) to Eq. (21)]:

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (18)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (19)$$

$$mAP = \frac{1}{n}\sum_{1}^{n} P(R)d(R) \quad (20)$$

$$F1 = \frac{2\times Precision\times Recall}{Precision+Recall} \times 100\% \quad (21)$$

where, TP (True Positives), FP (False Positives), and FN (False Negatives) represent the number of true positives, false positives, and false negatives, respectively. In combination, "TP + FP" is the total number of detected targets, and "TP + FN" is the total number of real targets in the image. mAP@0.5 represents the average value of mAP when the IoU threshold is 0.5. Among them, mAP@0.5:0.95 represents the average value of mAP under different IoU thresholds (from 0.5 to 0.95, with a step size of 0.05). Since the dataset has three category labels for different growth stages, n=3.

V. Results and Analysis

A. Comprehensive Evaluation of Model Performance

To evaluate the superiority of the Lite-CropNet model, the Lite-CropNet model was compared with five advanced object detection methods in terms of detection performance, including CenterNet [41], Faster R-CNN [24], EfficientDet [42], FCOS [43], and DETR [23]. The same training set and test set were used to train and test these models, and hyperparameter tuning was performed for all methods to ensure fair and objective results. The comprehensive evaluation indicators of each model for the three growth stages in the object detection task are shown in Table II.

TABLE II. Comparison of Evaluation Metrics of Different Models

Model	P	R	F1	mAP@0.5	mAP@0.5:0.95
CenterNet	73.5%	66.4%	69.8%	48.0%	15.9%
Faster R-CNN	57.5%	58.4%	57.9%	59.0%	25.4%
EfficientDet	52.1%	51.0%	51.5%	47.4%	16.8%
FCOS	62.4%	81.6%	70.7%	73.4%	44.7%
DETR	63.5%	72.7%	67.8%	61.8%	27.2%
YOLOv8	72.3%	81.6%	76.0%	83.3%	54.2%
YOLOv11	**77.0%**	75.1%	76.6%	77.1%	51.7%
Lite-CropNet	75.4%	**86.2%**	**80.4%**	**85.7%**	**56.8%**

The experiments show that the comprehensive performance of Lite-CropNet is superior to other object detection methods. The FCOS model ranks second in comprehensive performance,

with relatively high accuracy and recall for object detection. However, due to its single and lightweight network structure, it may have certain difficulties in handling occluded tomatoes and tomatoes of different sizes. CenterNet adopts an object detection method based on center points. It may have poor overall performance because the model focuses too much on the target center point and ignores other possible target parts. The Faster R-CNN and EfficientDet models show balanced performance in the P and R indicators, but their overall performance is not excellent. It is worth noting that the DETR model is different from other object detection models. It adopts an end-to-end training method, which reduces cumbersome steps and manual intervention. However, it is easily limited by annotated data and computing resources, and has high computational costs, which is not suitable for real-time application scenarios, and does not achieve good performance in this task.

In summary, the Lite-CropNet model performs best in the tomato detection task. The advantages of Lite-CropNet may come from its strong adaptability to multi-scale, multi-growth stage, and occluded targets, as well as the appropriate feature decoding and optimization processing of the feature maps of the Encoder in the Decoder.

B. Lightweight Analysis of Different Models

When deploying the tomato detection and classification model to low-end edge devices, lightweight performance is a key factor. To comprehensively consider the lightweight performance of the model, two indicators—Frame Per Second (FPS) and model parameter quantity (Params)—were used to evaluate the lightweight effect of different models. The FPS indicator measures the number of frames that the model can process per unit time; a higher FPS value indicates that the model can process more input data in the same time and has higher real-time performance. Params reflects the number of model parameters; a smaller number of parameters means that the model is more lightweight, requires fewer computing resources during inference, and is directly related to the storage space of the model and the efficiency of computing resource usage. Their calculation methods are as follows [see Eq. (22) and Eq. (23)]:

$$FPS = \frac{1000}{pre-process+inference+NMS} \quad (22)$$

$$Params = i \cdot (k \cdot k) \cdot o + o \quad (23)$$

In the equations, is the input size, is the convolution kernel size, and is the output size. Among them, $pre-process$, $inference$, and Non-Maximum Suppression (NMS) represent the time required for preprocessing, inference, and non-maximum suppression of each image, respectively.

This experiment was conducted on a PC equipped with a relatively low-end Nvidia GTX 1650 GPU. Fig. 4 lists the comparison of lightweight parameters of different models. From the comparison results, the Lite-CropNet model achieves an excellent comprehensive level, with a high frame rate of 76.9 and lightweight parameters of only 4.4M. In contrast, although the EfficientDet model has a small number of parameters (15.0M), its frame rate is only 17.2, which is relatively low. The Faster R-CNN and FCOS models also have a large number of parameters. The DETR model has a frame rate of 21.3 and a parameter quantity of 158M; although it adopts an end-to-end training method, its performance is still relatively general here. The CenterNet model has a high frame rate, but its parameter quantity is still not sufficient for deploying lightweight models. Although YOLOv11 shows balanced performance in accuracy (mAP@0.5 is 77.1%) and frame rate (63.6FPS), its parameter quantity (6.3M) and computational complexity are still higher than those of Lite-CropNet. This further proves that the advantages of Lite-CropNet in comprehensive performance come from its lightweight design, providing a more feasible solution for real-time applications on low-end devices. Especially in the agricultural field where available resources are limited, lightweight models are usually more favored because they require fewer computing resources, which is important for some agricultural managers and can effectively reduce their economic burden.

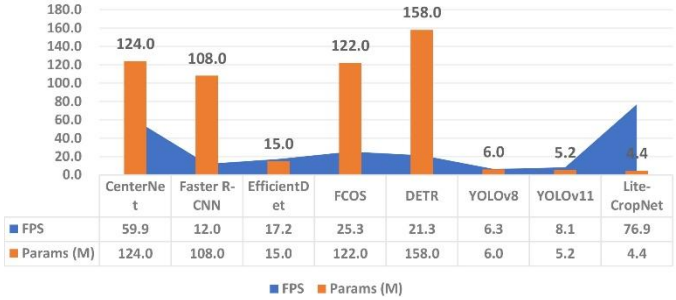


Fig. 4. FPS and Params comparison of different models.

C. Inference Speed Analysis of Different Models in a CPU Environment

In agricultural scenarios, many low-end edge devices (such as small embedded controllers and low-cost industrial computers) may not be equipped with GPUs and only rely on CPUs for inference. Therefore, a supplementary test on the inference speed of mainstream non-YOLO series models in the CPU environment was conducted. The experimental equipment was an Intel Celeron J4125 processor (4 cores and 4 threads, main frequency 2.0GHz) commonly used in agricultural scenarios, with 8GB of memory, and the test image resolution was still 640×640. The experimental data are shown in Table III.

TABLE III. COMPARISON OF EVALUATION METRICS OF DIFFERENT MODELS IN A GPU OR CPU ENVIRONMENT

Model	CPU FPS	GPU FPS (GTX 1650)	Parameter Count (M)
Lite-CropNet	**8.2**	**76.9**	**4.4**
CenterNet	6.5	59.9	108.0
Faster R-CNN	2.1	25.3	124.0
FCOS	3.3	21.3	122.0
DETR	1.8	17.2	158.0
EfficientDet	4.7	12.0	15.0

The results show that Lite-CropNet can still reach 8.2FPS in the CPU environment, which meets the basic requirements of real-time agricultural monitoring (usually requiring ≥5FPS). This performance benefits from its lightweight design:

compared with Faster R-CNN (2.1FPS) and DETR (1.8FPS), the parameter scale of Lite-CropNet is only 3.5% and 2.8% of the former, respectively, reducing memory occupation and computational delay. Compared with the same-level EfficientDet (4.7FPS), the CPU inference speed of Lite-CropNet is increased by 74%, which is due to the more concise feature fusion strategy adopted by its decoder, which reduces computational complexity. The Lite-CropNet in this study further optimizes the attention mechanism and loss function, and while maintaining accuracy, it achieves more efficient CPU inference, providing a feasible solution for agricultural scenarios without GPUs.

D. Visualization Analysis of Inference Results

By comparing the model prediction results with the real annotation results, the reasons affecting the detection performance of the model can be analyzed. Fig. 5 shows typical error cases, where the red arrows point to the regions missed by the model. Compared with other models, this model fails to correctly mark the real target regions. Missed detection is defined as failing to detect fruits with an occluded area of less than 50%, which is the standard for annotating the training set and test set in the TomatOD dataset.

From the experiment, it was observed that false detections mainly occur due to similar appearances, where leaves or debris are mistakenly identified as tomatoes. However, it is reassuring that in the test set, except for the case where the occlusion degree of the region pointed by the red arrow in the figure exceeds and is close to 50%, no missed detections were found in other images. Compared with the baseline model YOLOv11, Lite-CropNet achieves a lower missed detection rate in occluded scenarios (such as leaf occlusion and fruit overlap) and low-light environments. This indicates that the model can better capture tomato fruits under various environmental conditions, and even in dim environments with insufficient light, the model still shows similar discrimination ability to humans, further proving that the model has strong robustness and generalization.

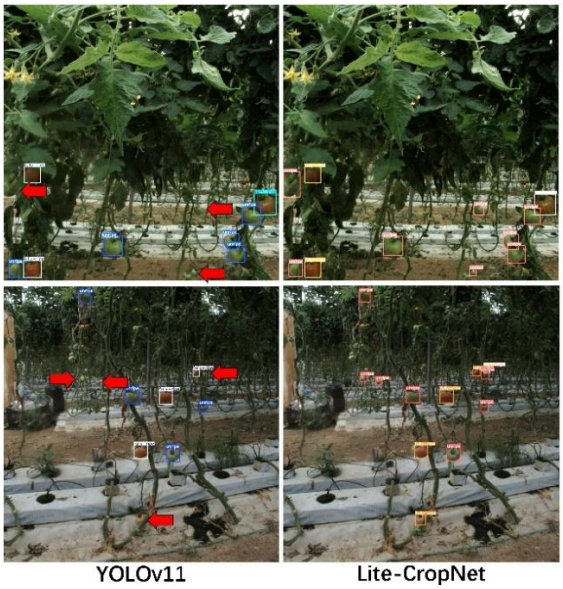


Fig. 5. Inference error results of the baseline model YOLOv11 and the Lite-CropNet model. Red arrows indicate missed detections.

E. Ablation Experiments

1) Ablation experiment on attention mechanism: In the above content, common attention mechanism methods are listed. To verify the detection differences of the Lite-CropNet model for tomatoes at each growth stage after adding different attention mechanisms to the upsampling and downsampling modules, ablation experiments were conducted using different attention mechanisms to analyze their performance differences.

The experimental results are shown in Table IV. The use of the Shuffle Attention mechanism has a positive impact on the model performance. In particular, the unique channel shuffling design of Shuffle Attention makes the mAP@0.5 indicator of the model at least 1.6% higher than that of other attention mechanisms. It can better capture global and local features, thereby comprehensively understanding the image content. Moreover, its mAP@0.5:0.95 ranks first at 57.4%, which is the only mechanism among all schemes that exceeds 87% in the core accuracy indicator. This design brings better-balanced performance in practical applications, enabling the model to achieve high levels in multiple key indicators.

TABLE IV. PERFORMANCE COMPARISON OF LITE-CROPNET USING DIFFERENT ATTENTION MECHANISMS

Attention	Precision (P)	Recall (R)	F1-Score (F1)	mAP@0.5	mAP@0.5:0.95
(None)	76.8%	80.7%	78.7%	85.5%	55.9%
SE[a]	78.4%	85.5%	**81.8%**	84.9%	56.9%
ECA[b]	80.2%	79.8%	80.0%	85.5%	56.2%
CBAM[c]	77.8%	79.3%	78.5%	83.8%	54.8%
Mlt_ECA[d]	75.4%	**86.2%**	80.4%	85.7%	56.8%
SA[e]	**80.4%**	79.5%	79.9%	**87.3%**	**57.4%**

[a] SE = Squeeze-and-Excitation

[b] ECA = Efficient Channel Attention

[c] CBAM = Convolutional Block Attention Module

[d] Mlt_ECA = Multi-scale Efficient Channel Attention

[e] SA = Shuffle Attention

In general, these observations indicate that the proposed Lite-CropNet model exhibits excellent robustness and generalization in practical tomato detection applications, providing a reliable automated solution for the agricultural field.

2) Ablation experiment on four-scale detection heads: To verify the adaptive value of the core innovation of Lite-CropNet—the "four-scale detection heads"—for crop targets of different sizes, this experiment constructs comparative models with only the number of detection heads modified. It focuses on analyzing the impact of the number of detection heads on the detection accuracy of small targets, large targets, and overall targets, so as to clarify the effectiveness of the scale design.

In this experiment, the core architecture of the model (CSPDarknet Encoder, Shuffle Attention, SIoU loss) was kept unchanged, and only the number of detection heads was adjusted: a two-scale model (2 detection heads, removing the

P2 small-scale and P5 large-scale branches) and a three-scale model (3 detection heads, removing the P5 large-scale branch) were constructed, and compared with the original four-scale model (4 detection heads, i.e., Lite-CropNet, retaining the full-scale branches of P2/P3/P4/P5). The performance comparison of the three groups of models on the same training set and validation set is shown in Table V.

TABLE V. PERFORMANCE COMPARISON OF MODELS WITH DIFFERENT NUMBERS OF DETECTION HEADS

Model Scale	P	R	F1	mAP@0.5	mAP@0.5:0.95
2	70.9%	82.2%	76.1%	78.2%	49.4%
3	72.7%	82.3%	77.2%	79.7%	50.7%
4 (Lite-CropNet)	**75.4%**	**86.2%**	**80.4%**	**85.7%**	**56.8%**

The experimental results show that the full-scale detection performance of the original four-scale model (4 detection heads) of Lite-CropNet is significantly better than that of the two-scale (2 detection heads) and three-scale (3 detection heads) comparative models: its overall detection accuracy (mAP@0.5 of 85.7%) and comprehensive performance index (F1 of 80.4%) are both at the optimal level. This proves that the four-scale detection heads can effectively improve the model's detection robustness for crop targets in the full scene by integrating feature information of different levels, including P2 (small-scale), P3 (medium-scale), P4 (medium-large scale), and P5 (large-scale), and reduce missed detections and false detections caused by insufficient scale adaptation. From the perspective of adaptive ability for segmented targets, the addition of the small-scale branch (P2) is the key to solving the problem of small fruit detection, which can effectively improve the model's recognition accuracy for small targets such as tomato young fruits and edge small fruits; while the large-scale branch (P5) can strengthen the positioning ability for large targets such as mature large tomatoes, avoiding bounding box deviation caused by the lack of global semantic information.

In summary, the four-scale detection heads enable Lite-CropNet to better adapt to multi-scale crops, solve the problem of insufficient detection ability of single-scale or few-scale models for targets of extreme sizes (such as extremely small young fruits and extra-large mature fruits), achieve a balance between accuracy and efficiency, and provide universal adaptive capabilities for multi-variety crop detection.

VI. DISCUSSION

In this research work, the Lite-CropNet model was proposed. Considering the cost of hardware resources, the lightweight design of the model was focused on. The TomatOD dataset was used to conduct in-depth research and performance analysis on the problem of crop fruit object detection and classification. Experiments have shown that while maintaining high accuracy, Lite-CropNet has a smaller model parameter size and higher frame rate, making it suitable for real-time application scenarios on low-end devices. In addition, the TomatOD dataset used was collected from real tomato planting scenarios, considering various challenges such as light changes and occlusions, resulting in a model with better robustness and generalization.

Furthermore, although the Lite-CropNet model generally performs well, the research still has potential shortcomings. First, a major problem is the relatively small scale of the dataset. It only contains 277 images and 2,413 tomato samples, and was collected from a single-region soilless cultivation greenhouse on Crete, Greece. It does not cover diverse planting environments (such as open fields and high-humidity solar greenhouses) in different countries and regions, nor does it include extreme scenarios such as rainstorm reflections and abnormal fruit appearances caused by diseases and pests, resulting in limited data diversity. Although the dataset already includes various environmental conditions and occlusion situations and successfully addresses them, the limited amount of data and the limitations of scene coverage may have a certain impact on the cross-region and cross-scene generalization ability and robustness of the model. Larger-scale datasets can usually better capture the diversity of the real world. As mentioned in the visual analysis above, the model inevitably has false detection problems when facing similar backgrounds, which is also considered to be caused by this reason. Second, for different tomato varieties in different countries and regions, the appearance characteristics of the plants are affected by various factors such as growth environment, soil conditions, and climate, and there may be slight differences in appearance, such as differences in fruit size and color depth, which may also affect the performance of the model. At the same time, although the study mentions that the model can be transferred to other horticultural crops such as potatoes and strawberries, it has not conducted verification for the morphological characteristics of these crops (such as clustered distribution of strawberries and underground fruiting of potatoes), and the multi-crop adaptability still needs further testing. This requires considering a wider and more diverse dataset during model training to ensure that the model has good adaptability to plant varieties in different countries and regions [44], [45].

VII. CONCLUSION

Efficient and accurate detection and counting of crop fruits have long been a challenging task. In this study, a novel Lite-CropNet model was proposed, with YOLOv11 as the baseline, aiming to address the key issues in horticultural crop detection and classification. A lightweight encoder was adopted and a concise yet efficient decoder was designed, which alleviates the problems of information loss and degradation. The exploration of different attention mechanisms revealed that the Shuffle Attention mechanism exerts a positive effect, helping to improve the model's performance. More importantly, efficiency enhancement was prioritized in the model design: a more lightweight architecture with a smaller parameter scale was developed, providing a cost-effective and highly efficient automated solution for agricultural production.

Experiments demonstrate that Lite-CropNet performs excellently under various lighting and occlusion conditions, exhibiting strong robustness and generalization. Notably, Lite-CropNet does not sacrifice efficiency at the expense of accuracy; instead, it achieves a balance between the two. Its

efficient performance in the CPU environment (8.2FPS) expands its application scenarios, making it particularly suitable for low-cost monitoring devices in greenhouses (e.g., Raspberry Pi-based embedded systems). In the future, CPU inference speed can be further improved through model quantization.

This work not only provides advanced technical support for the agricultural field but also offers insights for future research directions. The next phase of research will expand the dataset scale to cover more horticultural crop varieties, thereby enhancing the model's adaptability. Additionally, efforts will be made to further extend Lite-CropNet's performance advantages in the CPU environment and deploy it on low-cost monitoring devices in greenhouses, enabling it to better adapt to agricultural environments with extremely limited resources. Meanwhile, adversarial elements will be introduced to improve and optimize the object detection model, enhancing its robustness in complex scenarios. It is expected that this study will inspire more researchers to engage in agricultural intelligence and automation, and make more in-depth contributions to this field.

ACKNOWLEDGMENT

The research is supported by the Research Project of Software Engineering Institute of Guangzhou (No: KY202405).

REFERENCES

[1] F. Maureira, K. Rajagopalan, and C. O. Stöckle, "Evaluating tomato production in open-field and high-tech greenhouse systems," J. Cleaner Prod., vol. 337, pp. 130459, Apr. 2022.

[2] M. Afonso, H. Fonteijn, F. S. Fiorentin, D. Lensink, M. Mooij, N. Faber, R. Wehrens, "Tomato fruit detection and counting in greenhouses using deep learning," Front. Plant Sci., vol. 11, pp. 571299, Nov. 2020.

[3] L. Gong, M. Yu, S. Jiang, V. Cutsuridis, and S. Pearson, "Deep learning based prediction on greenhouse crop yield combined TCN and RNN," Sensors, vol. 21, no. 13, pp. 4537, Jul. 2021.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, May 2015.

[5] A. M. Hasan, F. Sohel, D. Diepeveen, H. Laga, and M. G. Jones, "A survey of deep learning techniques for weed detection from images," Comput. Electron. Agric., vol. 184, pp. 106067, May 2021.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Adv. Neural Inf. Process. Syst., vol. 25, pp. 1097-1105, 2012.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Las Vegas, NV, USA, 2016, pp. 770-778.

[8] F. Waldner and F. I. Diakogiannis, "Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network," Remote Sens. Environ., vol. 245, pp. 111741, Aug. 2020.

[9] J. Chen, J. Zhou, Q. Li, H. Li, Y. Xia, R. Jackson, J. Zhou, "CropQuant-Air: an AI-powered system to enable phenotypic analysis of yield- and performance-related traits using wheat canopy imagery collected by low-cost drones," Front. Plant Sci., vol. 14, pp. 1219983, Apr. 2023.

[10] V. Bathini and K. U. Rani, "A review of analyzing different agricultural crop yields using artificial intelligence," Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 1, pp. 1-10, 2025.

[11] V. Tsironis, S. Bourou, and C. Stentoumis, "tomatOD: Evaluation of object detection algorithms on a new real-world tomato dataset," in ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., vol. XLIII-B3-2020, pp. 1077-1084, 2020.

[12] S. Ahmad, Z. Chen, S. Ikram, and A. Ikram, "AI-enabled vision transformer for automated weed detection: Advancing innovation in agriculture," Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 12, pp. 1-9, 2024.

[13] Y. Zhang, Y. Xu, J. Hou, and Y. Song, "LMS-YOLO11n: A lightweight multi-scale weed detection model," Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 1, pp. 1-8, 2025.

[14] R. Reedha, E. Dericquebourg, R. Canals, and A. Hafiane, "Transformer neural network for weed and crop classification of high resolution UAV images," Remote Sens., vol. 14, no. 3, pp. 592, Feb. 2022.

[15] J. M. López-Correa, H. Moreno, A. Ribeiro, and D. Andújar, "Intelligent weed management based on object detection neural networks in tomato crops," Agronomy, vol. 12, no. 12, pp. 2953, Dec. 2022.

[16] Y. Mu, T. S. Chen, S. Ninomiya, and W. Guo, "Intact detection of highly occluded immature tomatoes on plants using deep learning techniques," Sensors, vol. 20, no. 10, pp. 2984, May 2020.

[17] F. Su, Y. Zhao, G. Wang, P. Liu, Y. Yan, and L. Zu, "Tomato maturity classification based on SE-YOLOv3-MobileNetV1 network under nature greenhouse environment," Agronomy, vol. 12, no. 7, pp. 1638, Jul. 2022.

[18] U. F. Rahim and H. Mineno, "Tomato flower detection and counting in greenhouses using faster region-based convolutional neural network," J. Image Graph., vol. 8, no. 4, pp. 107-113, Dec. 2020.

[19] U. F. Rahim, T. Utsumi, and H. Mineno, "Deep learning-based accurate grapevine inflorescence and flower quantification in unstructured vineyard images acquired using a mobile sensing platform," Comput. Electron. Agric., vol. 198, pp. 107088, Jul. 2022.

[20] H. K. Suh, J. Ijsselmuiden, J. W. Hofstee, and E. J. van Henten, "Transfer learning for the classification of sugar beet and volunteer potato under field conditions," Biosyst. Eng., vol. 174, pp. 50-65, Oct. 2018.

[21] D. K. Nkemelu, D. Omeiza, and N. Lubalo, "Deep convolutional neural network for plant seedlings classification," arXiv:1811.08404, Nov. 2018, unpublished.

[22] V. Tsironis, S. Bourou, and C. Stentoumis, "Tomatod: evaluation of object detection algorithms on a new real-world tomato dataset," Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., vol. 43, pp. 1077-1084, 2020.

[23] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in Proc. Eur. Conf. Comput. Vis., Glasgow, U.K., 2020, pp. 213-229.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," Adv. Neural Inf. Process. Syst., vol. 28, pp. 91-99, 2015.

[25] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," arXiv:2004.10934, Apr. 2020, unpublished.

[26] C. Y. Wang, H. Y. M. Liao, Y. H. Wu, P. Y. Chen, J. W. Hsieh, and I. H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops, Seattle, WA, USA, 2020, pp. 390-391.

[27] G. Jocher et al., "ultralytics/yolov5: v6.0 — YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support," Zenodo, Oct. 2021, doi: 10.5281/zenodo.5563715.

[28] Y. Sun, G. Chen, T. Zhou, Y. Zhang, and N. Liu, "Context-aware cross-level fusion network for camouflaged object detection," arXiv:2105.12555, May 2021, unpublished.

[29] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," Neural Netw., vol. 107, pp. 3-11, Nov. 2018.

[30] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in Proc. 14th Int. Conf. Artif. Intell. Stat., Fort Lauderdale, FL, USA, 2011, pp. 315-323.

[31] J. Xu, Z. Li, B. Du, M. Zhang, and J. Liu, "Reluplex made more practical: Leaky ReLU," in Proc. IEEE Symp. Comput. Commun., Rennes, France, 2020, pp. 1-7.

[32] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, 2018, pp. 7132-7141.

[33] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. Eur. Conf. Comput. Vis., Munich, Germany, 2018, pp. 3-19.

[34] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Seattle, WA, USA, 2020, pp. 11534-11542.

[35] Q. L. Zhang and Y. B. Yang, "SA-Net: Shuffle attention for deep convolutional neural networks," in Proc. IEEE ICASSP, Toronto, ON, Canada, 2021, pp. 2235-2239.

[36] Z. Yu, J. Ye, C. Li, H. Zhou, and X. Li, "TasselLFANet: A novel lightweight multi-branch feature aggregation neural network for high-throughput image-based maize tassels detection and counting," Front. Plant Sci., vol. 14, pp. 1158940, May 2023.

[37] Z. Gevorgyan, "SIoU loss: More powerful learning for bounding box regression," arXiv:2205.12740, May 2022, unpublished.

[38] D. Tzutalin, "LabelImg: Graphical image annotation tool," GitHub. https://github.com/tzutalin/labelImg, 2022.

[39] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," Adv. Neural Inf. Process. Syst., vol. 32, pp. 8024-8035, 2019.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980, Dec. 2014, unpublished.

[41] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in Proc. IEEE/CVF Int. Conf. Comput. Vis., Seoul, Korea, 2019, pp. 6569-6578.

[42] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Seattle, WA, USA, 2020, pp. 10781-10790.

[43] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in Proc. IEEE/CVF Int. Conf. Comput. Vis., Seoul, Korea, 2019, pp. 9627-9636.

[44] J. Ye, Z. Yu, Y. Wang, D. Lu, and H. Zhou, "WheatLFANet: In-field detection and counting of wheat heads with high-real-time global regression network," Plant Methods, vol. 19, no. 1, pp. 103, Dec. 2023.

[45] M. Gatto et al., "Trends in varietal diversity of main staple crops in Asia and Africa and implications for sustainable food systems," Front. Sustain. Food Syst., vol. 5, pp. 626714, Mar. 2021.

Using Combined Weighting and BP Neural Networks for Relative Poverty Measurement and its Evaluation

Xiaohua Cai[1], Ya Zhao[2]*, Lijia Chen[3], Juan Huang[4], Yang Xu[5]
School of International Business, Hunan University of Information Technology, Changsha 414000, China[1]
School of Management, Hunan University of Information Technology, Changsha 414000, China[2, 3, 4]
School of General Studies, Hunan University of Information Technology, Changsha 414000, China[5]

***Abstract*—This study addresses the challenges of measuring and evaluating relative poverty by introducing a comprehensive evaluation model based on the Analytic Hierarchy Process (AHP)-entropy method and BP neural networks. A multidimensional evaluation index system was constructed through expert consultation and literature review. The AHP-entropy method was then employed to determine the weights of the evaluation indicators, ensuring objectivity and scientific validity. Additionally, the BP neural network model was integrated to leverage self-learning and adaptive mechanisms for efficient and accurate poverty assessment. Empirical analysis shows that the model maintains a calculation error within 3.9%, demonstrating high precision and wide applicability. This research provides a novel approach that combines qualitative analysis with quantitative evaluation, offering a practical tool for governmental agencies to design effective poverty alleviation strategies. Moreover, the model opens new pathways for future research in regional poverty assessment, especially in enhancing cross-cultural adaptability and advancing intelligent evaluation models.**

***Keywords*—Analytic hierarchy process (AHP); entropy method; BP neural network model; relative poverty measurement**

I. Introduction

With the rapid development of the socioeconomic landscape, the issue of relative poverty has gradually become a crucial indicator for measuring social equity and harmony. Traditional poverty measurement methods, which predominantly rely on single income indicators, often fail to comprehensively capture the multidimensional characteristics of poverty. Therefore, research on poverty measurement in the context of the new era has shifted from a single income dimension to a comprehensive multidimensional assessment [1-2]. Multidimensional poverty encompasses not only economic income but also aspects such as education, health, and environment, reflecting a holistic understanding and in-depth analysis of poverty. How to achieve relative poverty measurement across multiple dimensions is an important topic in social science research, playing a significant role in achieving targeted poverty alleviation, promoting social equity, and improving people's living standards [3].

Past studies on relative poverty have often emphasized qualitative analysis [4-6], limiting their capacity for scientific quantitative assessments [7-8]. In response, some scholars have utilized the traditional Alkire-Foster (A-F) method. For example, Qin et al. [9] applied the A-F method to analyze multidimensional relative poverty in six regions of Xinjiang, while Gan et al. [10] combined the A-F method, Dagum Gini coefficient, and Kernel density estimation to examine multidimensional poverty and regional disparities across 31 provinces in China between 2013 and 2019. Similarly, Guo [11] employed CFPS data and a modified A-F method with subjective and objective weighting techniques to assess multidimensional poverty in China. Despite its widespread use, the A-F method has limitations, such as difficulties in precisely quantifying non-economic dimensions of poverty [12-13] and inherent subjectivity.

Regression analysis has also been explored as an alternative. Zhu et al. [14], for instance, employed logistic regression to identify factors influencing rural household poverty in western China, while Tian et al. [15] analyzed multidimensional poverty and its mechanisms in Henan Province using the A-F model and binary logistic regression. However, traditional mathematical approaches face challenges, including difficulties in constructing evaluation indicators, susceptibility to subjective bias, and computational complexity.

In contrast, the Analytic Hierarchy Process (AHP) offers significant advantages in quantifying and evaluating poverty. For instance, Dong [16] applied AHP and intergenerational income elasticity coefficients to explore factors affecting intergenerational poverty transmission. Luo and Cao [17] used AHP to construct a multidimensional indicator system for measuring urban household poverty in China. Zhu et al. [18] combined an AHP-improved A-F method to measure and decompose multidimensional relative poverty. Tao [19] established a multidimensional evaluation system with 14 indicators across five dimensions, using AHP and entropy-based weighting to determine indicator weights. Similarly, Chen [20] applied AHP to assess urban poverty during China's transition period, determining weights for various poverty dimensions.

Despite its utility, AHP has limitations, including subjectivity, high consistency requirements, potential complexity in hierarchical structures, challenges in determining weights, limited sensitivity analysis for parameter variations, constrained applicability, and significant computational demands.

To overcome these challenges, this study proposes a comprehensive evaluation model integrating AHP, the entropy method, and BP neural networks. The model begins by analyzing the factors influencing relative poverty and constructing a corresponding evaluation index system. Next,

*Corresponding author.

indicator weights are determined using AHP and the entropy method, which are then applied to score the relative poverty population. A BP neural network model is subsequently developed, trained, and tested using MATLAB. The BP neural network model proposed in this study provides more accurate evaluation results compared to traditional mathematical models and fuzzy mathematical theory. Through self-learning and adaptive adjustment, it effectively addresses the shortcomings of traditional evaluation methods in handling dimensional diversity and imprecise problem quantification.

By combining AHP and entropy-based weighting with BP neural networks, this study enhances the scientific validity and precision of poverty measurement while providing robust tools for poverty alleviation policy development and implementation. Continuous innovation and deeper research in this field can make meaningful contributions to reducing relative poverty, fostering social equity, and improving living standards.

II. METHODS

A. Construction of the Multidimensional Poverty Measurement, Evaluation System and Establishment of Data Samples

1) Construction of the measurement indicator system: The development of a multidimensional poverty measurement and evaluation system must adhere to the principles of scientific objectivity, comprehensiveness, feasibility, and representativeness [21]. For example, Huang et al. [22] classified poverty into seven dimensions: income standards, the "three guarantees" (education, healthcare, and housing), food and nutrition structure, social relations, and mental state. Similarly, Liao et al. [23] categorized poverty into dimensions such as living standards, production resources, income, health, education, political participation, sanitation facilities, and household assets. Thus, constructing a scientifically robust and rational evaluation indicator system is critical for addressing poverty measurement, with the design of these indicators being the core challenge.

At present, no unified standard exists for relative poverty evaluation systems. Therefore, this study develops its indicator system by interpreting the concept of relative poverty and utilizing statistical data from sources such as the China Statistical Yearbook and CFPS2020. Indicators with unavailable data were excluded. The finalized system reflects the practical context of China's economic development. It comprises five secondary indicators—basic living security, capabilities, social security, development opportunities, and economic conditions—and 14 tertiary indicators, as detailed in Table I.

2) Data sources: The quality of the sample used for learning plays a crucial role in determining the performance of the neural network model. Therefore, this study utilizes data from CFPS2020 (China Family Panel Studies). The data collection process adhered to rigorous scientific methods, tracking data across three levels: individual, household, and community. This approach ensured both the comprehensiveness and depth of the study. The broad geographic coverage, spanning 25 provinces, municipalities, and autonomous regions, further strengthened the generalizability and applicability of the findings. For data processing, STATA software, a widely recognized statistical tool, was employed to ensure accuracy and validity. During the processing, survey data were carefully filtered according to the dimensions of relative poverty measurement, resulting in 4,594 valid responses, thereby ensuring the study's precision and reliability.

3) Quantitative processing of the indicator system: Following data integration, the data were sequentially organized and transformed into the format shown in Table II, where each column represents the score of an individual respondent across 14 dimensions. A base score of 60 was assigned to each indicator, with specific values assigned according to the actual context, as outlined in Table II. This process finalized the quantitative analysis of each indicator.

TABLE I MULTIDIMENSIONAL RELATIVE POVERTY MEASUREMENT AND EVALUATION SYSTEM

Primary Indicators	Secondary Indicators	Tertiary Indicators
Multidimensional Relative Poverty Measurement and Evaluation System (A)	Living Security (W1)	Adequate Food and Clothing (W11)
		Drinking Water Supply (W12)
		Housing Conditions (W13)
	Economic Conditions (W2)	Per Capita Household Income (W21)
		Proportion of Transfer Income (W22)
	Feasible Capability (W3)	Education Level of Household Head (W31)
		Number of Laborers (W32)
		Health Status (W33)
	Development Opportunities (W4)	Access to Bank Loans (W41)
		Job Satisfaction (W42)
		Proportion of Migrant Laborers (W43)
	Social Security (W5)	Satisfaction with Medical Conditions (W51)
		Medical Insurance and Assistance (W52)
		Satisfaction with Education (W53)

B. Empowerment of the Multidimensional Poverty Measurement and Evaluation System

1) AHP-based empowerment of evaluation indicators: The Analytic Hierarchy Process (AHP) is a method that integrates both quantitative and qualitative analysis, enabling a comprehensive assessment of the research problem [18]. In this study, AHP is employed to determine the weight coefficients of each indicator. To minimize the randomness introduced by evaluators' subjective preferences and prior knowledge, a panel of experts was convened to provide a collective empowerment judgment. The expert panel constructed a criterion-level judgment matrix using the 1–9 scale method, as illustrated in Table III.

TABLE II QUANTITATIVE PROCESSING OF THE MULTIDIMENSIONAL RELATIVE POVERTY MEASUREMENT INDICATOR SYSTEM

Primary Indicators	Secondary Indicators	Tertiary Indicators	Tertiary Indicator Weights
Multidimensional Relative Poverty Measurement and Evaluation System (A)	Living Security (W1)	Adequate Food and Clothing (W11)	The total expenditure on food and clothing exceeds 2300, assign a value of 90; between 1380 and 2300, assign a value of 70; less than 1380, assign a value of 60.
		Drinking Water Supply (W12)	Tap water/bottled water/purified water/filtered water, assign a value of 90; well water, assign a value of 80; river/lake/pond spring water, assign a value of 70; others, assign a value of 60.
		Housing Conditions (W13)	Courtyards, villas, and townhouses, assign a value of 90; small buildings, assign a value of 80; unit apartments, assign a value of 70; single-story houses, assign a value of 60.
	Economic Conditions (W2)	Per Capita Household Income (W21)	Household per capita non-operating income below 50% of the median, assign a value of 60; between 50%–80% of the median, assign a value of 70; between 80%–100% of the median, assign a value of 80; above the median, assign a value of 90.
		Proportion of Transfer Income (W22)	Household savings below 50% of the median, assign a value of 60; between 50%–80% of the median, assign a value of 70; between 80%–100% of the median, assign a value of 80; above the median, assign a value of 90.
	Feasible Capability (W3)	Education Level of Household Head (W31)	Bachelor's degree or higher (including associate degree), assign a value of 90; high school/vocational school/technical school/secondary technical school, assign a value of 80; middle school, assign a value of 70; primary school or below, assign a value of 60.
		Number of Laborers (W32)	If the respondent's comprehension ability is weak and judged as very poor, assign a value of 60; otherwise, assign a value of 90.
		Health Status (W33)	If the respondent feels unhealthy, assign a value of 60; otherwise, assign a value of 90.
	Development Opportunities (W4)	Access to Bank Loans (W41)	If yes, assign a value of 60; if no, assign a value of 90.
		Job Satisfaction (W42)	Very satisfied, assign a value of 90; fairly satisfied, assign a value of 80; somewhat dissatisfied/average, assign a value of 70; very dissatisfied, assign a value of 60.
		Proportion of Migrant Laborers (W43)	If the respondent's information level about their future is below 50% of the median, assign a value of 60; between 50%–80% of the median, assign a value of 70; between 80%–100% of the median, assign a value of 80; above the median, assign a value of 90.
	Social Security (W5)	Satisfaction with Medical Conditions (W51)	Very satisfied, assign a value of 90; fairly satisfied, assign a value of 80; somewhat dissatisfied/average, assign a value of 70; very dissatisfied, assign a value of 60.
		Medical Insurance and Assistance (W52)	None, assign a value of 60; if yes, assign a value of 90.
		Satisfaction with Education (W53)	Very satisfied, assign a value of 90; fairly satisfied, assign a value of 80; somewhat dissatisfied/average, assign a value of 70; very dissatisfied, assign a value of 60.

TABLE III CRITERION-LEVEL JUDGMENT MATRIX

A	W1	W2	W3	W4	W5	Weight
W1	1	2	4	4	2	0.36
W2	0.5	1	5	4	2	0.28
W3	0.25	0.2	1	2	0.25	0.08
W4	0.25	0.25	0.5	1	0.2	0.06
W5	0.5	0.5	4	5	1	0.22

The matrix is solved, yielding a maximum eigenvalue of $\lambda_{Max} = 5.23$. The corresponding eigenvector is $W_i = (0.36, 0.28, 0.08, 0.06, 0.22)$. A consistency test is then conducted on W_i, as shown in Eq. (1):

$$CI = \frac{\lambda_{Max} - n}{n-1} = 0.0585 \quad (1)$$

According to the table, when n=5, the value of RI=1.12. Therefore, the relative consistency index can be calculated, as shown in Eq. (2):

$$CR = \frac{CI}{CR} = 0.05226 < 0.1 \quad (2)$$

The calculations demonstrate that the matrix achieves satisfactory consistency, meaning that the weight vector W_i is both objective and acceptable. Similarly, applying the same method to solve the other matrices results in the judgment matrices presented in Table IV to Table VII.

Table IV presents the weight distribution for the multidimensional relative poverty measurement indicator system. The weight of the secondary indicators represents their share within the overall evaluation system, while the weight of the tertiary indicators reflects their share within the corresponding secondary indicators. The comprehensive weight is the weighted average of both secondary and tertiary indicator weights, indicating their relative importance within the entire evaluation framework.

TABLE IV HP-BASED WEIGHT DISTRIBUTION FOR THE MULTIDIMENSIONAL RELATIVE POVERTY MEASUREMENT AND EVALUATION SYSTEM

Primary Indicators	Secondary Indicators	Secondary Indicator Weights	Tertiary Indicators	Tertiary Indicator Weights	Comprehensive Weight
Multidimensional Relative Poverty Measurement and Evaluation System (A)	Living Security (W1)	0.36	Adequate Food and Clothing (W11)	0.1	0.04
			Drinking Water Supply (W12)	0.23	0.08
			Housing Conditions (W13)	0.67	0.24
	Economic Conditions (W2)	0.28	Per Capita Household Income (W21)	0.8	0.22
			Proportion of Transfer Income (W22)	0.2	0.06
	Feasible Capability (W3)	0.08	Education Level of Household Head (W31)	0.09	0.01
			Number of Laborers (W32)	0.27	0.02
			Health Status (W33)	0.64	0.05
	Development Opportunities (W4)	0.06	Access to Bank Loans (W41)	0.12	0.01
			Job Satisfaction (W42)	0.68	0.04
			Proportion of Migrant Laborers (W43)	0.2	0.01
	Social Security (W5)	0.22	Satisfaction with Medical Conditions (W51)	0.26	0.06
			Medical Insurance and Assistance (W52)	0.63	0.14
			Satisfaction with Education (W53)	0.11	0.04

2) Entropy method for weight assignment to evaluation indicators: The entropy method is an objective weighting approach that assigns weights based on the degree of variation in indicator values. Its main advantage is that it avoids the subjective biases inherent in the Analytic Hierarchy Process (AHP), thereby improving both the objectivity and accuracy of weight assignment. This method is particularly effective for situations where indicators exhibit high variability, randomness, and disorder, providing a more accurate representation of each indicator's true significance within the evaluation system. Consequently, this study employs the entropy method to objectively assign weights to the evaluation indicators, following the steps outlined below:

a) Let A denote the j-th indicator of the i-th household. The corresponding attribute decision matrix is as shown in Eq. (3):

$$M = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{n1} \\ x_{21} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \quad (3)$$

b) Data standardization is performed to eliminate the influence of different units on the evaluation indicators. In this study, the mean normalization method is applied for standardization, and the calculation is as shown in Eq. (4):

$$y_{ij} = \frac{x_{ij}}{\overline{x_j}} \quad (4)$$

c) The contribution of the *i*-th sample under the *j*-th indicator, denoted as p_{ij}, is calculated as shown in Eq. (5):

$$p_{ij} = \frac{y_{ij}}{\sum_{i=1}^{m} y_{ij}} \quad (5)$$

d) The total contribution of the *j*-th indicator is calculated as its information entropy value, as shown in Eq. (6):

$$E_j = -k \sum_{j}^{m} p_{ij} ln(p_{ij}) \quad (6)$$

In the equation, $k = \frac{1}{\ln(m)}$.

e) Calculate the coefficient of variation for the contributions of each sample under the *j*-th indicator, as shown in Eq. (7):

$$d_j = 1 - E_j \quad (7)$$

*f)*Calculate the weight corresponding to the *j*-th indicator, as shown in Eq. (8):

$$w_j = \frac{d_j}{\sum_{j=1}^{m} d_j} \quad (8)$$

The weights calculated using the entropy method, based on the data from Section II B and analyzed with the R software, are summarized in the Table V.

3) Comprehensive weighting approach integrating AHP and the entropy method: The Analytic Hierarchy Process (AHP) is highly subjective, relying heavily on qualitative components and limited quantitative data, which reduces its credibility. In contrast, the entropy method lacks the ability to compare indicators across dimensions. To address these limitations, a combined weighting method is employed, which considers the intrinsic statistical properties and authoritative values of the indicators. This method integrates subjective and objective weighting approaches, balancing their respective strengths and weaknesses. The corresponding equation is as shown in Eq. (9):

$$W_j = \frac{\sqrt{\alpha_j \beta_j}}{\sum_{j=1}^{n} \sqrt{\alpha_j \beta_j}} \quad (9)$$

Among these, α_j denotes the weight values derived from the Analytic Hierarchy Process (AHP), while β_j represents the weight values obtained using the entropy method.

The combined weights are calculated using the combined weighting formula, based on the subjective and objective weight values obtained from the Analytic Hierarchy Process (AHP) and the entropy method. The corresponding weighting results are presented in Table VI.

TABLE V WEIGHT DISTRIBUTION FOR THE MULTIDIMENSIONAL RELATIVE POVERTY MEASUREMENT AND EVALUATION SYSTEM BASED ON THE ENTROPY METHOD

Primary Indicators	Secondary Indicators	Tertiary Indicators	Entropy Weight
Multidimensional Relative Poverty Measurement and Evaluation System (A)	Living Security (W1)	Adequate Food and Clothing (W11)	0.08
		Drinking Water Supply (W12)	0.02
		Housing Conditions (W13)	0.08
	Economic Conditions (W2)	Per Capita Household Income (W21)	0.02
		Proportion of Transfer Income (W22)	0.04
	Feasible Capability (W3)	Education Level of Household Head (W31)	0.12
		Number of Laborers (W32)	0.04
		Health Status (W33)	0.06
	Development Opportunities (W4)	Access to Bank Loans (W41)	0.09
		Job Satisfaction (W42)	0.05
		Proportion of Migrant Laborers (W43)	0.06
	Social Security (W5)	Satisfaction with Medical Conditions (W51)	0.03
		Medical Insurance and Assistance (W52)	0.24
		Satisfaction with Education (W53)	0.06

TABLE VI COMPREHENSIVE WEIGHT DISTRIBUTION FOR THE MULTIDIMENSIONAL RELATIVE POVERTY MEASUREMENT AND EVALUATION SYSTEM BASED ON AHP & ENTROPY METHOD

Primary Indicators	Secondary Indicators	Tertiary Indicators	Subjective Weight	Objective Weight	Comprehensive Weight
Multidimensional Relative Poverty Measurement and Evaluation System (A)	Living Security (W1)	Adequate Food and Clothing (W11)	0.04	0.08	0.07
		Drinking Water Supply (W12)	0.08	0.02	0.05
		Housing Conditions (W13)	0.24	0.08	0.16
	Economic Conditions (W2)	Per Capita Household Income (W21)	0.22	0.02	0.08
		Proportion of Transfer Income (W22)	0.06	0.04	0.06
	Feasible Capability (W3)	Education Level of Household Head (W31)	0.01	0.12	0.04
		Number of Laborers (W32)	0.02	0.04	0.03
		Health Status (W33)	0.05	0.06	0.07
	Development Opportunities (W4)	Access to Bank Loans (W41)	0.01	0.09	0.04
		Job Satisfaction (W42)	0.04	0.05	0.05
		Proportion of Migrant Laborers (W43)	0.01	0.06	0.03
	Social Security (W5)	Satisfaction with Medical Conditions (W51)	0.06	0.03	0.05
		Medical Insurance and Assistance (W52)	0.14	0.24	0.22
		Satisfaction with Education (W53)	0.04	0.06	0.06

C. Development of the BP Neural Network Model

1) Determining the fundamental parameters of the BP neural network: To construct a BP neural network, key parameters such as the number of layers, the number of neurons, the activation function, and the learning rate must be determined. Generally, increasing the number of layers enhances computational accuracy; however, this also results in longer training times and may increase the risk of overfitting. Alternatively, computational accuracy can be improved by adding more hidden nodes without expanding the network layers. Consequently, this study employs a three-layer BP neural network with a single hidden layer.

The number of neurons in each layer is determined as follows: the 14 tertiary indicators from the multidimensional relative poverty measurement system serve as input neurons, while the system's poverty scores are used as output neurons, resulting in $n = 14$ and $l = 1$. Based on Eq. (10), the range of values for the hidden layer nodes is [4, 8]. Therefore, this study sets the number of hidden layer nodes m to the higher value of 8 to enhance computational accuracy.

$$\sqrt{nl} \le m \le \sqrt{n(l+3)} + 1 \qquad (10)$$

In this study, the activation function for the input layer is defined as the hyperbolic tangent function, as shown in Eq. (11):

$$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \quad (11)$$

The activation function for the output layer is the Sigmoid function, as shown in Eq. (12):

$$f(x) = \frac{1}{1+e^{-x}} \quad (12)$$

The learning rate governs the adjustment of synaptic values during the neural network training process. If the learning rate is too high or too low, it can lead to system instability. To maintain stability during training, the learning rate (denoted as η) should be within the range of 0.01 to 0.1. In this study, η is set to 0.55. Based on these considerations, the topology of the BP neural network used in this study is shown in Fig. 1.

2) Parameter settings for the BP neural network model: To train the BP neural network model, the initial network parameters must be determined. These parameters typically include the weights for each layer and the activation thresholds. The weights and thresholds are initialized to small random values. In this study, the initial values are randomly selected within the range of [-1/12, 1/12] prior to training. Additionally, 70% of the data is designated as the training sample, with $x(m) = [x_{m1}, \ldots x_{m16}]$ representing the input data for the input layer, y_m representing the desired output, and m representing the sample index.

Once the initial parameters and training samples are set, the outputs of the hidden and output layers can be calculated. The induced local domain for a particular neuron in the hidden layer, as used in this study, is shown in Eq. (13):

$$v_j(m) = \sum_{i=1}^{n} w_{ij} x_{ij} - \theta_j \quad (13)$$

The induced local domain of a specific neuron k in the output layer is shown in Eq. (14):

$$v_k(m) = \sum_{j=1}^{s} w_{jk} v_j(m) - \theta_k \quad (14)$$

In the equation, n and s denote the number of neurons in the input and hidden layers, respectively.

Based on the induced local domain outlined above, the output of neuron j in the hidden layer at time m is as shown in Eq. (15):

$$y_j(m) = \frac{e^{v_j(m)}-e^{-v_j(m)}}{e^{v_j(m)}+e^{-v_j(m)}} \quad (15)$$

At this stage, the output of neuron k in the output layer is as shown in Eq. (16):

$$y_k(m) = \frac{1}{1+e^{-v_k(m)}} \quad (16)$$

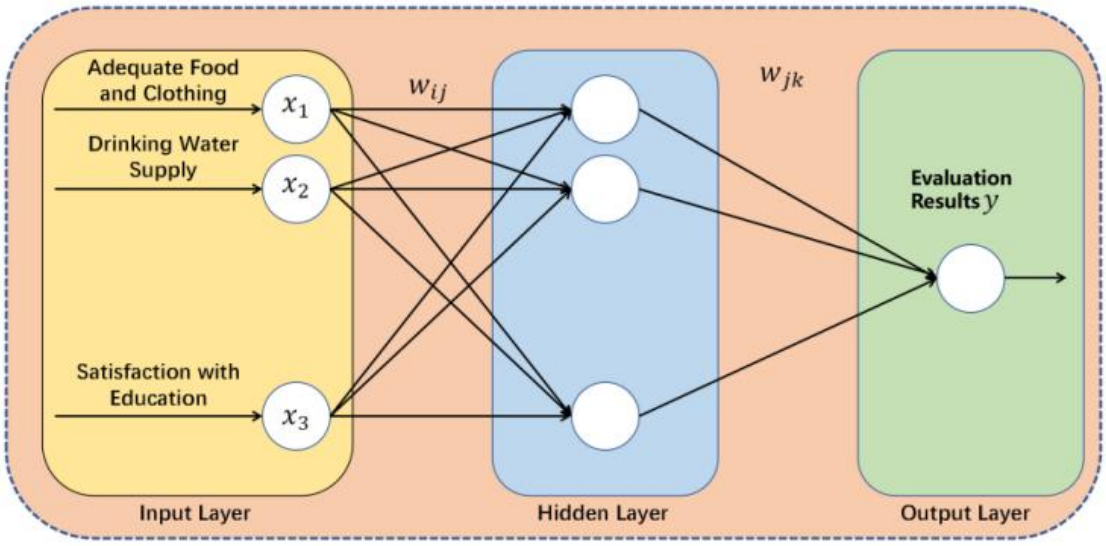


Fig. 1. Multidimensional relative poverty measurement and evaluation system model.

III. Results and Discussion

A. Model Training

After constructing the model as outlined above, 70% of the integrated data was used for training, with 15% allocated as training samples and the remaining 15% as testing and validation data. Fig. 2 illustrates the regression of the target relative to the output. The R-values for the training set, validation set, test set, and all data are 0.98371, 0.985, 0.97912, and 0.98313, respectively. These results suggest that the neural network model demonstrates a high degree of linearity and exhibits strong fitting performance.

B. Empirical Analysis

After training the neural network model, 20 sets of integrated data were selected for testing. The test results, compared with the comprehensive weighted calculations, are presented in Table VII (excerpt). The comparison reveals that the average relative error for the 20 test samples is 0.89%, with a maximum relative error of 3.90%. These results suggest that the model achieves high prediction accuracy. Furthermore, the use of the neural network model for multidimensional relative poverty measurement is simple, significantly reducing computational costs and demonstrating strong feasibility. The fitted curve is shown in Fig. 3.

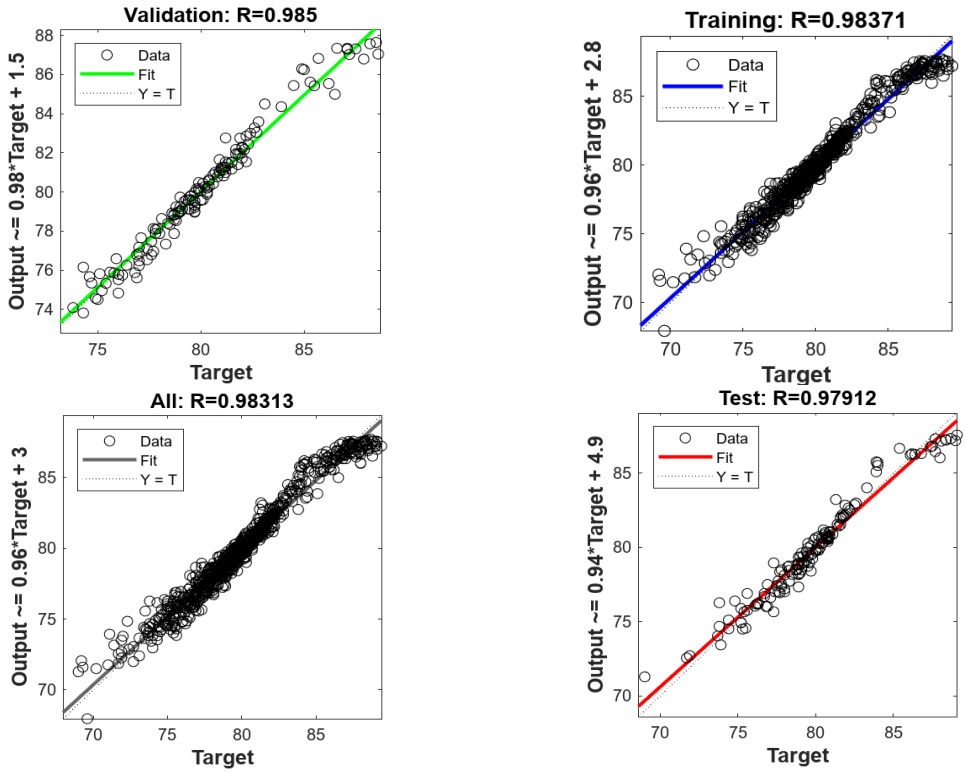


Fig. 2. Regression of the target relative to the output.

TABLE VII BP NEURAL NETWORK CALCULATION RESULTS AND COMPREHENSIVE WEIGHTING EVALUATION RESULTS

Serial Number	Expected Output	BP Predicted Value	Relative Error	Serial Number	Expected Output	BP Predicted Value	Relative Error
1	86.1	86.8574	0.88%	11	84.4	84.6768	0.33%
2	72.7	72.5812	0.16%	12	71.9	73.6172	2.39%
3	78.7	78.5607	0.18%	13	80.4	80.5731	0.22%
4	73.8	73.3003	0.68%	14	80.7	79.4826	1.51%
5	85.6	87.1798	1.85%	15	79.5	79.1354	0.46%
6	74.7	74.0416	0.88%	16	74.5	74.6252	0.17%
7	74.5	72.9669	2.06%	17	87.8	88.1500	0.40%
8	72	69.1954	3.90%	18	76.5	76.2431	0.34%
9	76.7	76.4779	0.29%	19	81.1	80.9983	0.13%
10	77.3	77.0135	0.37%	20	78.4	77.8524	0.70%

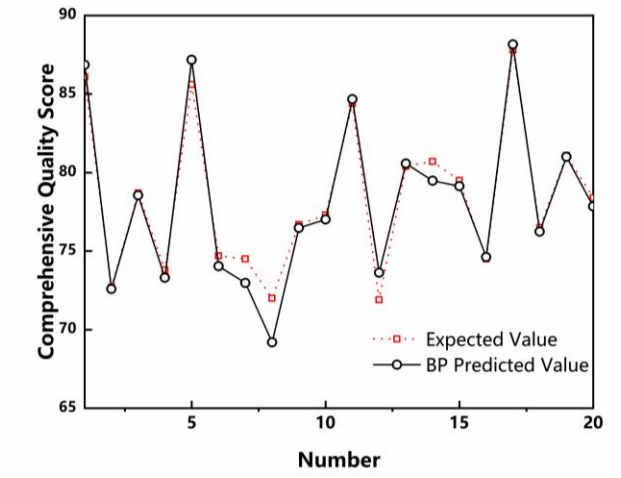


Fig. 3. Comparison curve of expected output and BP predicted value.

IV. CONCLUSIONS

This study develops a multidimensional relative poverty measurement evaluation system based on the AHP and entropy weight method, offering a quantitative analysis-based evaluation approach. To enhance the convenience of comprehensive quality assessment, a model combining the AHP and entropy weight method with the BP neural network is also proposed. The main conclusions are as follows:

This study employs the AHP-entropy method to allocate comprehensive weights to evaluation indicators, establishing a scientific and rational multidimensional relative poverty assessment system. The development of this framework provides governments with a quantitative and systematic evaluation tool during the implementation of multidimensional poverty assessment, thereby enhancing the objectivity and scientific rigor of the evaluation process.

The research demonstrates that the BP neural network-based relative poverty measurement model exhibits significant advantages in addressing complex nonlinear problems. The application of this model not only improves evaluation efficiency and reduces computational costs but also validates its feasibility and effectiveness in practical poverty population assessment through empirical analysis.

The BP neural network model proposed in this study holds broad application prospects in the field of poverty population assessment. This methodology enables more objective and equitable evaluation of multidimensional relative poverty, facilitating better implementation of poverty eradication policies by government departments and promoting comprehensive national development.

CONFLICTS OF INTEREST

The authors declare no potential conflict of interest.

REFERENCES

[1] Zhang WJ, Cheng SR, Xiu WY, et al. A comparative study on the multidimensional relative poverty differences in agricultural and pastoral areas under the goal of common prosperity—based on the multidimensional relative poverty measurement of 810 household survey data. Chinese Agricultural Resources and Zoning: 1-11.

[2] Liu L. Dynamic identification and measurement of multidimensional relative poverty of farmers under the background of common prosperity. Agricultural Economics, 2023(12): 96-97.

[3] Ruan J, Wang JT, Yang X. From "Lacking" to "Unequal"—Research on relative poverty measurement from the perspective of shared prosperity. Mathematical Statistics and Management: 1-12.

[4] Tian YJ. Research and application of measurement methods for the effectiveness of targeted poverty alleviation. Capital University of Economics and Business, 2021.

[5] Rong SS. Research on rural poverty measurement, causality analysis, and fiscal poverty alleviation. Southwestern University of Finance and Economics, 2022.

[6] Hong Y. Measurement and causality analysis of inter-provincial rural household energy poverty in China. Zhejiang University of Finance and Economics, 2024.

[7] Huang XY, Xu ML, Chen ML. Study on multidimensional relative poverty measurement and its influencing factors—a case study of the interprovincial border area of Hunan, Hubei, Chongqing, and Guizhou. Contemporary Rural Finance and Economics, 2023(11): 2-7.

[8] An JQ, Li Q, Xu YD. Measurement and governance of relative poverty in rural China in the Post-Poverty Era. Agricultural Outlook, 2023, 19(02): 67-72.

[9] Qin HB, Li WH, Xie ZX. Multidimensional measurement and analysis of relative poverty in Xinjiang—Based on surveys of six regions (Prefectures/ Cities) in Northern and Southern Xinjiang. Arid Zone Resources and Environment, 2023, 37(02): 16-25.

[10] Gan XC, Cai YY, Xiao HB. Multidimensional relative poverty measurement and its distribution dynamics in China. Statistics and Decision, 2023, 39(06): 50-55.

[11] Guo XH. Measurement of rural multidimensional poverty based on the improved alkire-foster model — A case study of City A. Northeast Agricultural Science, 2024, 49(01): 98-104.

[12] Xue DX. Measurement and analysis of multidimensional poverty in China and its influencing factors. Liaoning University, 2022.

[13] Liu Z, Xie YM, Ding S. Gray relational quantitative evaluation method and multidimensional relative poverty measurement. Statistics and Decision, 2023, 39(21): 40-45.

[14] Zhu L, Li WZ, Liao HP, et al. Study on the measurement and influencing factors of rural relative poverty in Western China under the background of rural revitalization — Based on an empirical study of X Province. Southwest University Journal (Natural Science Edition), 2023, 45(04): 14-27.

[15] Tian MJ, Ma Z, Wang W, et al. Measurement of multidimensional poverty and poverty mechanisms in central agricultural areas — A case study of Xincai County in Henan Province. Regional Research and Development, 2023, 42(02): 143-149.

[16] Dong PM. Research on the intergenerational transmission and interruption path of multidimensional poverty. Shanghai University of Engineering Science, 2021.

[17] Luo XL, Cao YC. Proposal and empirical analysis of a measurement indicator system for urban family poverty in China Based on the AHP Method. Journal of Central University of Finance and Economics, 2010(06): 75-80.

[18] Zhu ZC, Xue DX, Sun CT. Multidimensional relative poverty measurement and decomposition based on the improved AHP-AF Method. Statistics and Decision, 2021, 37(16): 10-14.

[19] Tao J. Research on the Measurement of rural relative poverty from a multidimensional poverty perspective. Jiangxi University of Finance and Economics, 2021.

[20] Chen LZ. Study on urban poverty measurement during China's transition period. Huazhong University of Science and Technology, 2009.

[21] Yang F, Zhuang TH. Construction and empirical measurement of a multidimensional relative poverty identification indicator system for new generation migrant workers from the perspective of feasible capability. China's Western Region, 2021(06): 1-14.

[22] Huang Q, Liu W, Chen J. Identification and measurement of rural multidimensional poverty in the New Era. Statistics and Decision, 2022, 38(20): 16-22.

[23] Liao LZ, Yin Y, Chen SY. Study on multidimensional poverty measurement for farmers in Yunnan's 'Three Areas'. Southwest Forestry University Journal (Social Sciences Edition), 2019, 3(01): 23-26.

Evaluating Transformer-Based Pretrained Models for Classical Arabic Named Entity Recognition

Mariam Muhammed, Shahira Azab
Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt

***Abstract*—This study presents a comprehensive comparative evaluation of transformer-based pretrained language models for Named Entity Recognition (NER) in Classical Arabic, an underexplored linguistic variety characterized by rich morphology, orthographic ambiguity, and the absence of diacritics. The main objective of this work is to identify the most effective transformer model for Classical Arabic NER and to analyze the linguistic factors influencing model performance. Using the CANERCorpus, which contains Hadith texts annotated with twenty fine-grained entity types, ten transformer-based models were fine-tuned and evaluated under consistent experimental settings. The study benchmarks models such as AraBERT, ArBERT, and multiple CAMeLBERT variants, comparing their precision, recall, and F1-scores. The results demonstrate that all models achieve strong performance (F1 > 96%), while CAMeL-CA-NER attains the highest score (F1 = 97.78%), confirming the advantage of domain-specific pretraining on Classical Arabic data. Error analysis further reveals that domain-adapted models better handle ambiguous entities and religious terminology. A comparative analysis with traditional and non-transformer approaches, including rule-based and BERT-CRF models from previous studies, shows that CAMeL-CA-NER surpasses earlier methods by more than 3% in F1-score, highlighting its superior capability in handling Classical Arabic text. However, this study is limited to the CANERCorpus, which primarily consists of Hadith texts; results may vary for other Classical Arabic genres or domains. These findings provide a valuable benchmark for future research and demonstrate the adaptability of modern NLP architectures to linguistically complex, low-resource domains.**

***Keywords*—Classical Arabic; Named Entity Recognition; transformer models; pretrained models; CANERCorpus**

I. Introduction

Named Entity Recognition (NER) is an important task in natural language processing (NLP) that aims to identify and categorize entities such as persons, locations, and organizations. Since its introduction in the Message Understanding Conferences (MUC) in the 1990s [1], NER has become essential for various NLP applications, including information retrieval, machine translation, question answering, and text summarization [2], [3], [4].

Extensive research has been conducted in English and other widely spoken languages, including Spanish and Chinese. NER systems are generally built using three approaches: rule-based, machine learning, and hybrid methods. Rule-based systems rely on handcrafted linguistic rules and dictionaries [1], [5], [6], but they are limited by language-specific expertise and lack flexibility [7], [8]. Deep learning, a subset of machine learning, introduced neural architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which learn representations at both the character and word levels [9].

The rapid growth of Arabic content on the internet has created a strong need for accurate NLP tools for Arabic. Arabic serves as the official language in the Arab World, covering 22 countries [2]. It is a Semitic language with rich vocabulary, complex morphology, and challenging syntax. It exists in three forms, including Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA). CA represents the original Arabic language from the seventh to the early eleventh century CE [10]. MSA is used in formal contexts such as books, formal communications, and news articles, while DA is spoken in everyday life. Although CA is less used in modern contexts, it remains highly important as the language of the Quran (the holy book of Islam), the Hadith (the sayings of Prophet Muhammad), and the Islamic heritage.

Arabic NER faces unique challenges compared to other languages due to its morphological richness, allowing suffixes and prefixes to function as conjunctions, prepositions, and pronouns, as in the word "فأسقيناكموه" (faskenakomoh), meaning "So we gave you water to drink". Unlike languages that rely on capitalization to identify proper nouns, Arabic does not use capitalization. Moreover, Arabic utilizes short vowels (diacritics) to disambiguate word sense, but it is usually written without them, leading to ambiguity in word meaning. For example, the word "درس" can mean "دَرْس" (lesson) as a noun or "دَرَسَ" (has studied) as a verb [6]. Some examples of ambiguous words in Arabic text are listed in Appendix A.

Classical Arabic, in particular, poses additional linguistic challenges beyond those of Modern Standard Arabic. In addition to its rich morphology, absence of diacritics, and orthographic ambiguity, Classical Arabic exhibits archaic vocabulary, complex syntactic structures, and context-dependent expressions that differ significantly from Modern Standard Arabic usage. These characteristics increase word-level ambiguity and make NER in Classical Arabic notably more difficult than in contemporary Arabic varieties.

Furthermore, the scarcity of datasets and other reliable resources for Arabic increases the challenges of Arabic NER [2], [11].

Recent advances in deep learning and pretrained models have significantly improved Arabic NER. Contextualized embeddings such as ELMo, BERT, and GPT, along with sequence-to-sequence and convolutional neural network models, have enhanced system performance by capturing the

complex linguistic features of Arabic and learning effectively from annotated data. Transformer-based pretrained language models have transformed NLP and had a major impact on Arabic NER. Trained on massive text corpora, they can identify complex linguistic patterns. When fine-tuned on Arabic NER datasets, these models achieve state-of-the-art results with less data and lower computational cost.

The main contribution of this study is a comparative evaluation of transformer-based pretrained models for Classical Arabic NER using the CANERCorpus dataset. It investigates their predictive capabilities and performance on Hadith texts. This study provides a systematic benchmarking of transformer-based pretrained models for Classical Arabic NER. Unlike previous studies that mainly target Modern Standard Arabic, this work focuses on Classical Arabic, analyzing model behavior in a linguistically rich and underexplored domain. This benchmarking effort serves as a foundation for future research aiming to develop specialized models and resources for Classical Arabic.

The remainder of this study is organized as follows: Section II reviews related work on Arabic NER, covering rule-based, traditional machine learning, and deep learning approaches, and highlights gaps in previous studies. Section III presents the pretrained transformer models used in this study, including their architectures, training data, and fine-tuning procedures for NER. Section IV describes the dataset, the experimental setup, and evaluation measures. Section V reports the experimental results, providing quantitative performance comparisons across models. Section VI offers a discussion of the results, analyzing model behavior and domain-specific challenges. Section VII presents a qualitative error analysis, highlighting common mistakes and ambiguous cases. Section VIII provides a comparative evaluation with previous Arabic NER approaches, emphasizing the benefits of domain-specific pretraining. Section IX discusses the limitations of the study, including the dataset and methodological constraints. Finally, Section X concludes the study and outlines potential directions for future research. Supplementary material, such as lists of ambiguous words and extended experiment details, is provided in Appendix A.

II. Related Work

Many approaches have been proposed to perform Arabic NER. Qu et al. (2024) classified them into four main paradigms: rule-based methods, machine learning, deep learning, and pretrained language models [10].

Before the rise of machine learning, the rule-based approach was the prime choice utilized in NER tasks. These systems depended on manually written linguistic rules and dictionaries. Examples of features commonly used include morphological analyzers [8], [12], [13], lexical triggers [14], regular expressions and gazetteers [15], as well as transliteration techniques [16]. Rule-based systems usually perform well in narrow domains but fail to generalize across domains. They also require heavy manual effort, which makes them less scalable [1].

With the advancement of statistical approaches, machine learning (ML) became widely adopted. Researchers explored several ML techniques, including conditional random fields (CRF) [17], [18], support vector machines (SVM) [19], [20], and meta-classifiers [21], [22], [23]. These approaches combined features such as lexical, contextual, morphological, gazetteer, and part-of-speech tags to improve performance. Although machine learning methods achieve satisfactory performance, they struggle to learn complex and high-level features from data when using linear models such as log-linear HMM or linear chain CRF. However, these approaches often underperform on Classical Arabic due to the language's rich morphology, absence of diacritics, and the scarcity of domain-specific datasets.

To address this limitation, deep learning methods have emerged as a solution for automatically discovering hidden features. The advancement of Arabic NER through deep-learning methods can be analyzed from three perspectives: i) input representations such as word-level embeddings [22], character-level embeddings [24], and additional features ; ii) context encoders, such as CNN [25], RNN [24], [26], or multi-attention networks [26] that capture dependencies between tokens; and iii) label decoders that assign the correct entity labels [25], [27]. While these methods reduce reliance on manual feature engineering, they still struggle to disambiguate entities in Classical Arabic texts, especially in religious and historical domains.

In recent years, pretrained language models (PLMs) have transformed Arabic NER. By learning contextual representations from massive corpora, models such as BERT [27], AraBERT [28], ARBERT and MARBERT [29], ArabicBERT [30], CAMelBERT [31], and Multilingual BERT [32] achieved state-of-the-art performance in many NLP tasks, including NER [33]. Despite their success, most studies focus on Modern Standard Arabic or social media text. Classical Arabic, particularly in Hadith texts, poses unique challenges not fully addressed by existing models.

Our study builds on this prior work by applying transformer-based pretrained models specifically to Classical Arabic NER using the CANERCorpus. This approach leverages domain-specific pretraining to improve recognition of ambiguous words, religious entities, and morphologically complex tokens, addressing gaps left by previous approaches.

III. Pretrained Models

We conducted training on the dataset using multiple pretrained language models. Table I provides an overview of the used models and their training data.

- AraBERT: Developed by Antoun et al. [28], AraBERT is a BERT-based model specialized for Arabic. It has several versions (v0.1/v1 and v0.2/ v2). Versions v0.1 and v1 were trained on 23GB of text, while v0.2 and v2 used a larger 77GB corpus. In this study, we employed AraBERT (v2).
- ARBERT/MARBERT: Introduced by Abdul-Mageed et al. [29], these models are based on the BERT architecture. ARBERT was trained on 66GB of news text, while MARBERT was trained on a larger 128GB

dataset that included 50% tweets. MARBERT is particularly strong for social media Arabic.

- ARBERT/MARBERT: Introduced by Abdul-Mageed et al. [29], these models are based on the BERT architecture. ARBERT was trained on 66GB of news text, while MARBERT was trained on a larger 128GB dataset that included 50% tweets. MARBERT is particularly strong for social media Arabic.
- ArabicBERT: Developed by Safaya et al. [30], this model was trained on 95GB of Arabic text, mostly from the OSCAR corpus. It captures broad Arabic patterns and structures.
- GigaBERT: Developed by Lan et al. [34], GigaBERT is a bilingual model trained on English and Arabic. It was trained on ACE2005 data, including broadcast news and newsgroups.
- QARiB: Provided by Abdelali et al. [35], QARiB was trained on a huge dataset consisting of around 420 million tweets gathered from Twitter and 180 million text sentences gathered from Arabic GigaWord, Abulkhair, and OPUS corpora.
- CAMeLBERT Models: Developed by Inoue et al. [31], CAMeLBERT has three variants: CAMeLBERT-MSA-NER, CAMeLBERT-CA-NER, and CAMeLBERT-MIX-NER. They are fine-tuned for NER tasks on MSA, CA, and a mix of MSA/CA/DA texts.
- Multilingual BERT: Proposed by Devlin et al. [27], this model supports 104 languages, including Arabic. Although trained on Wikipedia across languages rather than Arabic-specific data, it still provides useful performance for Arabic NER.

TABLE I. OVERVIEW OF THE PRETRAINED MODELS

Model Name	Model Information		
	Authors	***Training Data***	***Arabic Text Source***
AraBERT (v2)	Antoun et al. [28]	77 GB	OSCAR, Arabic Wikipedia, OSIAN, Arabic Corpus
ARBERT	Abdul-Mageed et al. [29]	66 GB	News articles
MAR-BERT	Abdul-Mageed et al. [29]	128 GB	Tweets, News
Arabic-BERT	Safaya et al.[30]	95 GB	OSCAR corpus
Giga-Bert:	Lan et al. [34]	ACE-2005	Broadcast news, Newsgroups
QARIB	Abdelali et al. [35]	420M tweets + 180M sentences	Twitter API, GigaWord, Abulkhair, OPUS
CAMeLBERT-MSA	Inoue et al. [31]	107 GB	Gigaword, Wikipedia, OSCAR, OSIAN
CAMeLBERT-CA	Inoue et al. [31]	6 GB	OpenITI corpus (v1.2)
CAMeLBERT-MIX	Inoue et a. Mach	167 GB	Various Sourses Combined (MSA, CA, DA)
Multi-lingual BERT	Devlin et al. [27]	~16 GB total (all languages), Arabic ≈ 3 GB	Wikipedia (104 languages)

IV. EXPERIMENTS

In this section, we describe the dataset, experimental setup, and evaluation metrics used in this study.

A. Dataset

Table II presents a list of available annotated datasets for Arabic NER. In this study, we employed the "CANERCorpus", which will be further detailed in this section.

TABLE II. LIST OF ANNOTATED DATASETS FOR ARABIC NER

Corpus	Year	Text Source	Tags	words	Availability
ANERcorp[1]	2007	Website, news, magazines	4	150 k	Free
ACE 2003[2]	2004	Broadcast News, newswire genres	7	42 k	Required fees
ACE 2004[3]	2004	Broadcast News, newswire genres	7	151 k	Required fees
ACE 2005[4]	2005	Transcripts, news	7	~300 k	Required fees
ACE 2007[5]	2007	Transcripts, news	7	~300 k	Required fees
REFLEX[6]	2009	Reuters news	4	22.5 k	Required fees
AQMAR[7]	2012	Arabic Wikipedia	7	1 M	Free
OntoNotes 5.0[8]	2013	News	18	300 k	Required fees
WDC[9]	2014	Wikipedia	4	6 M	Free
CANER[10]	**2018**	**Religion**	**20**	**258 k**	**Free**
Wojood[11]	2022	Wikipedia	21	550 K	Under request

We used the CANERCorpus [36], a Classical Arabic NER dataset annotated by experts. It contains 7,000 Hadiths from *Sahih Al-Bukhari*. The corpus includes 20 entity classes such as person, location, organization, date, book, and others. In total, the dataset has about 258K words, with 72K entities (≈23%). Fig. 1 and Table III show word count for each Named Entity.

[1] URL: https://camel.abudhabi.nyu.edu/anercorp/
[2] URL: https://catalog.ldc.upenn.edu/LDC2004T09
[3] URL: https://catalog.ldc.upenn.edu/LDC2005T09
[4] URL: https://catalog.ldc.upenn.edu/LDC2006T06
[5] URL: https://catalog.ldc.upenn.edu/LDC2014T18
[6] URL: https://catalog.ldc.upenn.edu/LDC2009T11
[7] URL: http://www.cs.cmu.edu/~ark/ArabicNER/
[8] URL: https://catalog.ldc.upenn.edu/LDC2013T19
[9] URL: https://github.com/Maha-J-Althobaiti/Arabic_NER_Wiki-Corpus
[10] URL: https://github.com/RamziSalah/Classical-Arabic-Named-Entity Recognition-Corpus
[11] URL: https://ontology.birzeit.edu/Wojood/

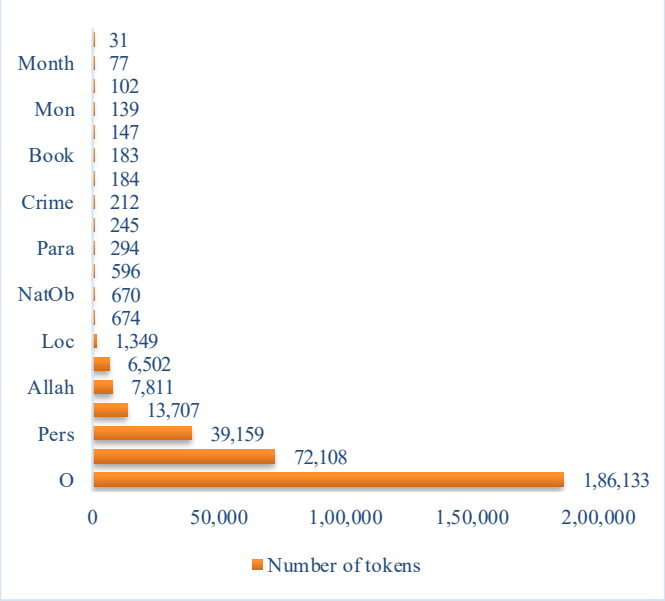


Fig. 1. Distribution of Named Entity tokens in CANERCorpus.

TABLE III. WORD COUNT FOR EACH NAMED ENTITY CLASS

Named Entity tag	Number of tokens
Allah	7,811
Prophet (Pro.)	6,502
Pers	39,159
Num	13,707
Loc	1,349
Clan	674
NatOb	670
Crime	212
Date	596
Para	294
Hell	245
Rlig	184
Book	183
Means	147
Mon	139
Time	102
Month	77
Day	31
Sect	17
Org	9
Other Words (O)	186,133
Named Entity	72,108
Total	258,241

For our experiments, we focused on person, God, prophet, location, clan, date, natural object, and other entities. The "number" entities were excluded since most of them were page numbers and did not add value to the task.

The CANERCorpus dataset did not provide clear sentence boundaries. We estimated the average sentence length in Sahih Al-Bukhari (38 words) and used this to split the text into sentences as in [36]. To avoid splitting related entities, the last token of each segment was labeled as "other" (O). We also removed punctuation, special characters, and diacritics. The dataset was then divided into training, validation, and test subsets. We randomly split the corpus into 80% training, 10% validation, and 10% testing. Table IV presents the distribution of named entities across each subset.

TABLE IV. THE DISTRIBUTION OF NAMED ENTITIES ACROSS THE TRAINING, TESTING, AND VALIDATION DATASETS

Named Entity	Training	Testing	Validation	Total
Pers	31,362	3,754	4,043	39,159
Allah	6,269	765	777	7,811
Prophet	5,162	669	671	6,502
Loc	1,087	127	135	1,349
Clan	601	103	63	767
NatOb	550	61	59	670
Date	491	56	49	596
Other	158,119	20,288	19,296	197,703
Total	203,641	25,823	25,093	254,557

B. Experimental Setup

The experiments were conducted on the Google Colab platform12 using a Tesla T4 GPU. Models were trained on the training dataset, with hyperparameters tuned on the validation dataset. Table V summarizes the parameters.

TABLE V. HYPERPARAMETER VALUES FOR THE MODELS

Parameter	Value
Learning Rate (LR)	1e-5
Maximum Length Size	256
Batch size	16
Optimizer	Adam
No. of epochs	Varied by model

C. Evaluation Measures

To assess the effectiveness of the pretrained models, we employed various metrics, including precision, recall, and F-measure. Eq. (1) outlines precision (P), which measures the correct identification of named entities by the model among all identified entities. Eq. (2) illustrates recall (R), indicating the correct identification of named entities by the model among all entities present in the corpus. Eq. (3) describes the F-measure (F), employed to harmonize the conflicting relationship between precision and recall [9], [37].

$$precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$F1-Score = \frac{2\ (Precicion*Recall)}{(precsion+Recall)} \quad (3)$$

where, TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

[12] https://colab.research.google.com/

V. RESULTS

In this section, we present the results of training ten pretrained models on the CANERCorpus dataset. Table VI and Fig. 2 summarize the models' performance in terms of Precision, Recall, and F1-score.

TABLE VI. RESULTS ACHIEVED BY PRETRAINED LANGUAGE MODELS

Model Name	P	R	F	No. of Epochs
AraBERT	96.11	96.94	96.53	13
ArBERT	96.89	**98.09**	97.47	7
MARBERT	96.51	97.44	96.97	7
GigaBERT	96.93	97.50	97.21	11
Arabic-BERT	96.63	97.50	97.06	8
QARIB	96.31	96.70	96.50	7
Camel-msa-ner	96.85	97.75	97.30	8
Camel-ca-ner	**97.51**	98.05	**97.78**	9
Camel-mix-ner	96.94	97.84	97.39	8
Multilingual	96.11	96.94	96.53	11

The bold numbers mark the best results.

All models achieved strong results, with F1 scores ranging from 96.53% to 97.78%. The best performance was obtained by CAMeLBERT-CA-NER with an F1 score of 97.78%. QARiB also performed strongly, and models such as AraBERT, ArabicBERT, ARBERT, and MARBERT achieved high results with F1 scores above 96.5%. In contrast, Multilingual BERT obtained the lowest F1 score (96.53%).

These results demonstrate the robustness of transformer-based models in handling the linguistic challenges of Classical Arabic.

VI. DISCUSSION

This section provides an analysis and interpretation of the results presented in Section V.

The best performance of CAMeLBERT-CA-NER confirms the advantage of domain-specific pretraining on Classical Arabic texts, since the model was exposed to linguistic and stylistic patterns directly relevant to the Hadith corpus. This pretraining allowed CAMeLBERT-CA-NER to better recognize religious and historical entities (e.g., Prophet names, clans, and locations) and to handle morphological variations typical of Classical Arabic. In contrast, models trained primarily on Modern Standard Arabic or social media data showed slightly higher misclassification rates for such entities due to domain mismatch. CAMeLBERT-MIX-NER also achieved competitive performance, suggesting that multiple Arabic varieties (MSA, CA, and DA) can further improve generalization across heterogeneous texts.

QARiB performed strongly as well, benefiting from its large-scale training on diverse Twitter and formal Arabic corpora. This highlights that large amounts of heterogeneous data, even if not exclusively Classical Arabic, can still capture rich contextual representations transferable to CA NER tasks.

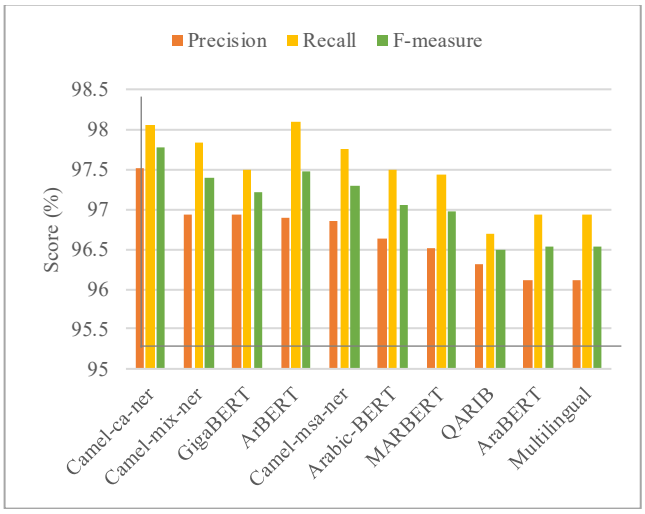


Fig. 2. Performance of pretrained models on CANERCorpus.

Models such as AraBERT, ArabicBERT, ARBERT, and MARBERT also achieved high results. Although primarily trained on MSA and social media text, they were able to generalize effectively to CA. This suggests that while surface differences exist between MSA and CA, pretrained models can leverage their shared morphological and syntactic features.

The weaker performance of Multilingual BERT can be attributed to the relatively small portion of Arabic in its training data (~3 GB from Wikipedia), compared to the much larger corpora used in Arabic-specific models. This result underlines the importance of language-specific and domain-focused training when dealing with morphologically rich languages like Arabic.

Beyond overall scores, a closer inspection of entity-level performance revealed interesting trends. Models trained on Classical Arabic data (e.g., CAMeLBERT-CA-NER) showed superior accuracy in recognizing religious and historical entities such as Prophet names, clans, and locations mentioned in Hadiths. On the other hand, models trained mainly on MSA and tweets occasionally misclassified such entities, reflecting a domain mismatch. However, these models performed comparably well on generic entities like dates, times, and numerical expressions, where the differences between CA and MSA are minimal.

Overall, the results highlight three important insights:

- Domain-specific pretraining matters. CAMeLBERT-CA-NER demonstrates clear advantages when the training domain matches the target text.
- Cross-variety transfer is feasible for MSA and DA-based models still generalize well to CA, especially for common entity types.
- Data size and diversity play a crucial role; larger and more varied corpora (e.g., QARiB, MARBERT) improve robustness, even if they are not domain-specific.

These findings emphasize that while transformer models are inherently powerful, their effectiveness for Classical Arabic NER is maximized when they are trained on domain-relevant corpora. At the same time, the strong performance of general-purpose models suggests promising opportunities for cross-domain transfer and low-resource adaptation in future research.

VII. Error Analysis

To better understand the behavior of the pretrained models on Classical Arabic NER, we conducted an analysis of detected ambiguities in the test set. We selected the four highest-performing models based on F1-score: ArBERT, CAMeL-CA-NER, CAMeL-MSA-NER, and CAMeL-MIX-NER. The test set contained 24,966 words. Table VII summarizes the number of correctly and incorrectly tagged words for these models.

TABLE VII. Correct and Incorrect Tag Counts for Top-Performing Models

Model	Correct Tags	Incorrect Tags
ArBERT	24,813	153
CAMeL-ca-NER	24,829	137
CAMeL-mas-NER	24,805	161
CAMeL-mix-NER	24,813	153

Some examples of ambiguous words and their corresponding model predictions are shown in Table VIII.

- Words following "ك" (e.g., كرسول) are sometimes misclassified by ArBERT, which labels them as "O" instead of the correct tag "Prophet".
- Words preceded by "ب", such as (*بنخلة*), may be incorrectly tagged as "LOC" by ArBERT and, in some cases, by CAMeL-CA-NER and CAMeL-MIX-NER, though the correct tag is "O".
- Words preceded by "ف", such as (*فلصفوان*) in the phrase (*وإن لم يرض عمر فلصفوان*), are occasionally misclassified by ArBERT as "O" instead of "Pers".
- Words following (*نو*) in CAMeL-CA-NER are sometimes wrongly labeled as "Pers", for example (*نو اليدين*).
- Some prophets' names, like (*ابراهيم*) and (*يوسف*), are occasionally tagged as "Pers" instead of "Prophet", and vice versa. This also occurs in ArBERT.
- Certain locations, for instance (*مزدلفة*) and (*جمرة العقبة*), are misclassified by CAMeL-CA-NER.
- Person names following words like (*امرأة*) or (*ابن*) may be incorrectly labeled as "Pers" even when they refer to prophets, as in (*ابن مريم*). This occurs in ArBERT, CAMeL-CA-NER, and CAMeL-MIX-NER.

Additional ambiguities observed specifically in ArBERT and CAMeL-MSA-NER include:

- Words like (*قوله*) followed by (*تعالى*) are misrecognized as "Allah" instead of "O".
- Nouns such as (*أسماء*) are sometimes tagged as "Pers" when they are not proper names, e.g., in (*أحب أسماء على رضي الله عنه إليه*).
- Words like (*على*) are occasionally misclassified as "Pers", e.g., (*وكانت على بردة*).
- Certain verbs followed by person names can be misclassified as "Pers", for example (*فنسي عوف ثم عمر بن الخطاب*).
- The word (*الله*) is sometimes labeled as "Allah" when it should be "*لنبي الله*" or "*لرسول الله*".

These errors mainly arise from differences between the models' training data and the Classical Arabic text in CANERCorpus. Models fine-tuned on Modern Standard Arabic (e.g., ArBERT, CAMeL-MSA-NER) show more ambiguities, while models trained on Classical Arabic (CAMeL-CA-NER) or a mixture of varieties (CAMeL-MIX-NER) exhibit fewer errors. The qualitative analysis in Table IX shows that CAMeL-CA-NER effectively handles Classical Arabic morphological patterns and religious expressions, such as prefixes (و, ب, ف) and embedded prophet names, that often confuse models trained solely on Modern Standard Arabic. Its superior handling of ambiguous cases like كرسول reflects its experience to Classical Arabic syntax and vocabulary during pretraining. In contrast, ArBERT and CAMeL-MSA-NER frequently misclassify tokens due to domain mismatch and limited exposure to archaic forms. Occasional over-generalization errors in CAMeL-CA-NER, such as mislabeling اليدين as Person, indicate sensitivity to morphological cues that resemble name patterns. These insights confirm that domain-relevant pretraining enhances contextual disambiguation and improves the recognition of specialized entity types in Classical Arabic texts.

TABLE VIII. Model Predictions for Ambiguous Words

The ambiguous word	Correct tag	AraBERT	CAMeL-ca-NER	CAMeL-msa-NER	CAMeL-mix-NER	Weighted- ET-ANER
وبمكة	Loc	O X	Loc √	Loc √	Loc √	Loc √
كرسول	Pro.	O X	Pro. √	Pro. √	Pro. √	Pro. √
اليدين	O	O √	Pers X	O √	O √	O √
قال فيوسف نبي الله	Pro.	O X	Pers X	Pro. √	Pro. √	Pro. √

VIII. Comparative Evaluation with Previous Arabic NER Approaches

Table IX and Fig. 3 compare the performance of the best model, CAMeL-CA-NER, with previously reported results on the CANERCorpus dataset. Early rule-based approaches achieved modest performance (F1 = 89.5%), reflecting the limitations of handcrafted rules for Classical Arabic. Later BERT-based models reached around 94 to 95% F1, demonstrating the benefits of contextualized embeddings. However, our CAMeL-CA-NER model achieved the highest score (F1 = 97.78%), outperforming earlier systems by more than 3%. This improvement can be attributed to its domain-specific pretraining on Classical Arabic texts (OpenITI corpus[13]), which provides a closer linguistic match to Hadith data. These findings confirm that transformer architectures with domain-relevant pretraining substantially advance the state of Classical Arabic NER.

TABLE IX. Comparison with Previous Works that Used CANERCorpus in Arabic NER Process

Previous Work	Approaches	Precision	Recall	F1-score
[38]	Rule-based approach	90.2	89.3	89.5
[39]	BERT	94.1	94.9	94.5
	BERT-CRF	94.0	95.4	94.7
	BERT-BLSTM-CRF	93.8	95.2	94.4
	BERT-GRU-CRF	94.1	95.5	94.8
[31]	CAMel-CA-NER	**97.51**	**98.05**	**97.78**

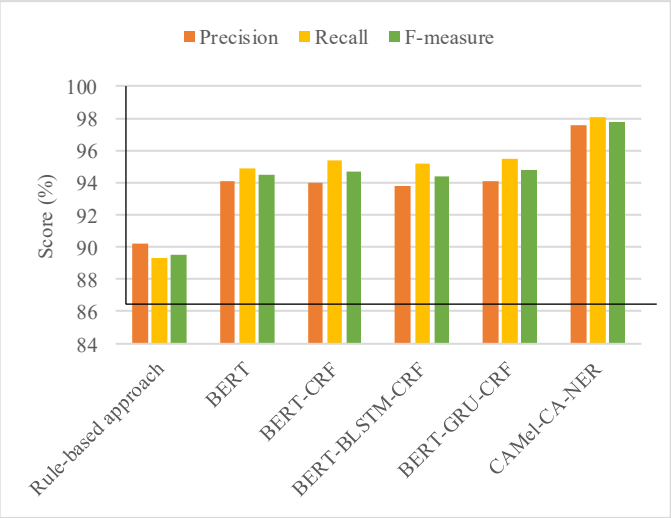


Fig. 3. Comparison of CAMel-CA-NER performance with previous Arabic NER approaches on the CANERCorpus dataset.

IX. Limitations of the Study

Despite the strong performance of transformer-based models for Classical Arabic NER demonstrated in this study, there are several limitations. First, the dataset used (CANERCorpus) is relatively outdated (2018), which may limit the applicability of our findings to more recent or varied Classical Arabic texts. Second, the corpus consists exclusively of Hadith texts, which means the models are primarily exposed to religious and historical language; this may affect their generalization to other Classical Arabic domains such as literature or formal documents. Third, while our models achieved high overall F1 scores, some entity types remain challenging, particularly ambiguous words or rare entities. Finally, the evaluation was limited to standard metrics (precision, recall, F1-score), and further work is needed to assess models' robustness in real-world NLP applications or cross-domain scenarios. Acknowledging these limitations provides transparency and helps readers interpret the scope of our findings accurately.

X. Conclusion and Future Work

This study evaluated the performance of transformer-based pretrained language models for Classical Arabic NER using the CANERCorpus of Hadith texts. The results showed that all models achieved strong performance, with F1 scores above 96%, confirming the effectiveness of transformer-based approaches for morphologically rich and linguistically complex languages. Among the models, CAMeLBERT-CA-NER achieved the highest F1 score (97.78%), demonstrating the advantages of domain-specific pretraining on Classical Arabic texts. While models trained on MSA and social media text generalized reasonably well to CA, the error analysis revealed that they were more prone to misclassifying religious and historical entities, emphasizing the importance of domain-relevant data for accurate NER. The novelty of this study lies in its systematic benchmarking and detailed linguistic analysis rather than proposing a new architecture. By systematically comparing multiple transformer-based models on Classical Arabic data, this work establishes a valuable reference point for future advancements in Arabic NER and provides insights for developing domain-specific pretrained models.

Future research can build on these findings in several directions. First, expanding the dataset with additional Classical Arabic sources, including underrepresented entity types, would enhance model coverage and reduce ambiguity-related errors. In particular, while the CANERCorpus (2018) remains a valuable resource, incorporating more recent datasets would improve the relevance and applicability of the findings. Second, investigating cross-domain transfer and domain adaptation between Classical Arabic, Modern Standard Arabic, and dialectal varieties could further improve adaptability and robustness across diverse text genres. Third, incorporating strategies to handle ambiguous words and diacritic variations may enhance model accuracy. Fourth, exploring model interpretability techniques to understand how transformer models capture linguistic cues and contextual dependencies would provide valuable insights into their decision-making process. Finally, integrating these pretrained models into downstream applications such as semantic search, question answering, and historical text analysis would demonstrate their practical utility and support broader NLP research in Classical Arabic.

[13] https://github.com/OpenITI/RELEASE/tree/v2019.1.1

REFERENCES

[1] A. Sharma, S. Chakraborty, and S. Kumar, "Named entity recognition in natural language processing: A systematic review," in Proceedings of Second Doctoral Symposium on Computational Intelligence, Springer, 2022, pp. 817–828.

[2] K. Darwish, N. Habash, M. Abbas, H. Al-Khalifa, H. T. Al-Natsheh, H. Bouamor, K. Bouzoubaa, V. Cavalli-Sforza, S. R. El-Beltagy, and W. El-Hajj, "A panoramic survey of natural language processing in the Arab world," Commun ACM, vol. 64, no. 4, pp. 72–81, 2021, doi: 10.1145/3447735.

[3] T. El Moussaoui and C. Loqman, "Advancements in Arabic Named Entity Recognition: A Comprehensive Review," IEEE Access, 2024.

[4] S. Albahli, "An Advanced Natural Language Processing Framework for Arabic Named Entity Recognition: A Novel Approach to Handling Morphological Richness and Nested Entities," Applied Sciences, vol. 15, no. 6, p. 3073, 2025.

[5] B. Jehangir, S. Radhakrishnan, and R. Agarwal, "A survey on named entity recognition—datasets, tools, and methodologies," Natural Language Processing Journal, vol. 3, p. 100017, 2023.

[6] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Named entity recognition and relation extraction: State-of-the-art," ACM Computing Surveys (CSUR), vol. 54, no. 1, pp. 1–39, 2021.

[7] D. N. Shah and H. B. Bhadka, "Named entity recognition from Gujarati text using rule-based approach," in Intelligent Systems Design and Applications: 17th International Conference on Intelligent Systems Design and Applications (ISDA 2017) held in Delhi, India, December 14-16, 2017, Springer, 2018, pp. 797–805.

[8] A. Anandika, S. Chakravarty, and B. K. Paikaray, "Named entity recognition in Odia language: a rule-based approach," International journal of reasoning-based intelligent systems, vol. 15, no. 1, pp. 15–21, 2023.

[9] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," IEEE Trans Knowl Data Eng, vol. 34, no. 1, pp. 50–70, 2020.

[10] X. Qu, Y. Gu, Q. Xia, Z. Li, Z. Wang, and B. Huai, "A Survey on Arabic Named Entity Recognition: Past, Recent Advances, and Future Trends," IEEE Trans Knowl Data Eng, vol. 36(03), pp. 943-959., 2024.

[11] M. Jarrar, M. Khalilia, and S. Ghanem, "Wojood: Nested arabic named entity corpus and recognition using bert," in Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 3626–3636.

[12] K. Shaalan, "Rule-based approach in Arabic natural language processing," The International Journal on Information and Communication Technologies (IJICT), vol. 3, no. 3, pp. 11–19, 2010.

[13] R. Alfred, L. C. Leong, C. K. On, and P. Anthony, "Malay named entity recognition based on rule-based approach," 2014.

[14] M. Hudhud, H. Abdelhaq, F. Mohsen, A. Rocha, L. Steels, and J. van den Herik, "ArabiaNer: A System to Extract Named Entities from Arabic Content.," in ICAART (1), 2021, pp. 489–497.

[15] K. Shaalan and H. Raza, "Person name entity recognition for Arabic," in Proceedings of the 2007 workshop on computational approaches to semitic languages: common issues and resources, 2007, pp. 17–24.

[16] D. Samy, A. Moreno, and J. M. Guirao, "A proposal for an Arabic named entity tagger leveraging a parallel corpus," in International Conference RANLP, Borovets, Bulgaria, 2005, pp. 459–465.

[17] M. A. Ali, A. B. A. Alwahhab, and Y. Farjami, "An Integrated Deep Learning Framework Combining LSTMCRF, GRU-CRF, and CNN-CRF with Word Embedding Techniques for Arabic Named Entity Recognition.," International Journal of Robotics & Control Systems, vol. 5, no. 2, 2025.

[18] M. S. Al-Qurishi and R. Souissi, "Arabic named entity recognition using transformer-based-crf model," in Proceedings of the 4th International Conference on Natural Language and Speech Processing (ICNLSP 2021), 2021, pp. 262–271.

[19] B. Ait Benali, S. Mihi, I. El Bazi, and N. Laachfoubi, "New approach for Arabic named entity recognition on social media based on feature selection using genetic algorithm," International Journal of Electrical and Computer Engineering, vol. 11, no. 2, p. 1485, 2021.

[20] R. M. Hamad and A. M. Abushaala, "Medical named entity recognition in arabic text using SVM," in 2023 IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA), IEEE, 2023, pp. 200–205.

[21] Y. Benajiba, I. Zitouni, M. Diab, and P. Rosso, "Arabic named entity recognition: using features extracted from noisy data," in Proceedings of the ACL 2010 conference short papers, 2010, pp. 281–285.

[22] M. Gridach, "Character-aware neural networks for arabic named entity recognition for social media," in Proceedings of the 6th workshop on South and Southeast Asian natural language processing (WSSANLP2016), 2016, pp. 23–32.

[23] I. Keraghel, S. Morbieu, and M. Nadif, "A survey on recent advances in named entity recognition," arXiv preprint arXiv:2401.10825, 2024.

[24] M. N. A. Ali, G. Tan, and A. Hussain, "Bidirectional recurrent neural network approach for Arabic named entity recognition," Future Internet, vol. 10, no. 12, p. 123, 2018.

[25] D. Awad, C. Sabty, M. Elmahdy, and S. Abdennadher, "Arabic name entity recognition using deep learning," in Statistical Language and Speech Processing: 6th International Conference, SLSP 2018, Mons, Belgium, October 15–16, 2018, Proceedings 6, Springer, 2018, pp. 105–116.

[26] M. N. A. Ali, G. Tan, and A. Hussain, "Boosting Arabic named-entity recognition with multi-attention layer," IEEE Access, vol. 7, pp. 46575–46582, 2019.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of NAACL-HLT, 2018, pp. 4171–4186.

[28] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," in Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 2020, pp. 9–15.

[29] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: deep bidirectional transformers for Arabic," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 7088–7105.

[30] 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, 2020, pp. 2054–2059.

[31] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, 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.

[32] S. Bilal, "A Linguistic System for Predicting Sentiment in Arabic Tweets," in 2021 3rd International Conference on Natural Language Processing (ICNLP), IEEE, 2021, pp. 134–138.

[33] B. Kessentini, T. Alimi, and R. Boujelben, "Assessing BERT Models for Arabic Named Entity Recognition," in Advancements in Machine Learning and Natural Language Processing: Innovations and Applications: 3rd International Conference on Language Processing and Knowledge Management (LPKM'2024), Springer Nature, 2025, p. 194.

[34] W. Lan, Y. Chen, W. Xu, and A. Ritter, "Gigabert: Zero-shot transfer learning from english to arabic," in Proceedings of The 2020 Conference on Empirical Methods on Natural Language Processing (EMNLP), 2020.

[35] A. Abdelali, S. Hassan, H. Mubarak, K. Darwish, and Y. Samih, "Pre-training bert on arabic tweets: Practical considerations," arXiv preprint arXiv:2102.10684, 2021.

[36] R. E. Salah and L. Q. B. Zakaria, "Building the classical Arabic named entity recognition corpus (CANERCorpus)," in 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), IEEE, 2018, pp. 1–8.

[37] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," Multimed Tools Appl, vol. 82, no. 3, pp. 3713–3744, 2023.

[38] R. Salah, M. Mukred, L. Qadri binti Zakaria, R. Ahmed, and H. Sari, "A new rule-based approach for classical arabic in natural language processing," Journal of Mathematics, vol. 2022, pp. 1–20, 2022.

[39] N. Alsaaran and M. Alrabiah, "Classical Arabic named entity recognition using variant deep neural network architectures and BERT," IEEE Access, vol. 9, pp. 91537–91547, 2021.

APPENDIX A: AMBIGUOUS ARABIC WORDS AND THE IMPACT OF DIACRITICS

Arabic form	English Meaning	Transliterations	Part of Speech (POS)
حُر	Freedom	Ḥurr	Noun
حَر	Hot	Ḥarr	Adjective
حَلْم	Patience	Ḥilm	Noun
حِلْم	Forgivingness	Ḥilm	Noun
حُلْم	Dream	Ḥulm	Noun
عِلْم	Knowledge	Elm	Noun
عَلَم	Flag	Alam	Noun
عَلِمَ	Knew	Alima	Verb
عُلِمَ	Is known	Ulima	Verb
عَلَّمَ	Taught	Allama	Verb
عُلِّمَ	Is taught	Ullima	Verb

A Hippopotamus Optimization Algorithm-Based Convolutional Neural Network Model for Mental Health Assessment Among College Students

Gai Hang[1], Lin Yang[2]*
President's Office of Guilin University of Aerospace Technology, Guilin, China[1]
School of Artificial Intelligence, Guilin University of Aerospace Technology, Guilin, China[2]

***Abstract*—The mental health of adult students is crucial not only for enhancing their learning experience and overall quality of life, but also for alleviating academic and employment-related anxiety. A significant challenge in developing effective online mental health support systems is the accurate assessment of students' mental health status. Current evaluation methods often lack precision and fail to integrate multifaceted data perspectives. To address these challenges, this study developed a psychological assessment system based on deep learning technology. The system aims to assess adult students' psychological states and provide appropriate support. Specifically, it utilizes a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) algorithm framework to evaluate students' psychological states by synthesizing image data, academic performance, and textual inputs. Furthermore, to enhance the accuracy of deep learning-based mental health assessment models, an improved hippopotamus optimization (IHO) algorithm was designed to optimize the hyperparameters of deep learning frameworks. By using the proposed multi-input single-output hybrid IHO-based LSTM-CNN framework (IHO-LSTM-CNN), the online mental health assessment module can accurately describe the psychological status of college students and provide personalized support to meet their specific needs. The final results indicate that the IHO-LSTM-CNN framework provides more accurate assessments than existing mental health assessment models, with an accuracy of 90.28%. This enhanced accuracy enables online community psychological support systems to deliver precise and effective psychological support to college students.**

***Keywords*—Convolutional Neural Network; Long Short-Term Memory; hippopotamus optimization algorithm; mental health assessment; deep learning**

I. Introduction

Amid growing global concern for mental health issues, researchers are designing generative artificial intelligence (AI)-based psychological support systems that utilize generative dialogue technology [1]. However, these systems often lack accuracy in assessing users' mental health status and consequently fail to provide high-quality, tailored psychological support [2]. Concurrently, higher education institutions and educational management departments are increasingly concerned about the rising prevalence of mental health problems among adult students. To address these specific challenges, an online community psychological support system powered by generative AI is designed. This system accurately assesses the mental health status of adult students based on their inquiry statements and basic information. Subsequently, leveraging the assessment results, it generates targeted psychological support statements to effectively address their mental health concerns.

The core goal of a generative AI-based network community psychological support system is to assess adult students' psychological states online, such as anxiety, depression, and stress levels, through non-invasive, continuous data collection and analysis. A core challenge for this online assessment system is accurately evaluating the psychological status of adult students based on available data. Specifically, since adult students' psychological data typically includes multimodal information (such as text, images, behavioral patterns, etc.), the evaluation model must effectively integrate these diverse inputs. Furthermore, the psychological state of adult students can fluctuate rapidly due to academic and family pressures, requiring algorithms to possess both short-term sensitivity and long-term trend prediction capabilities.

Current algorithms for mental health assessment primarily include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), traditional Random Forest algorithms, and expert rule-based systems [3]. Among these, CNN-based models excel at text sentiment analysis and image micro-expression recognition. However, CNN algorithms struggle to model dynamic changes in psychological states. They also demand large volumes of annotated data during processing, making the acquisition of psychological labels costly and subjective. In contrast, RNNs and Long Short-Term Memory Networks (LSTMs) can effectively capture the temporal evolution of psychological states. Nevertheless, their high model complexity can lead to overfitting, consequently reducing prediction accuracy. Finally, traditional Random Forest models are typically suitable only for small-sample static scenarios and struggle with unstructured data [4].

Another key challenge in generative AI-based network community psychological support systems is effectively using this technology to provide necessary psychological support for adult students. Current psychological support models primarily include those based on large language models (LLMs), knowledge graphs, and reinforcement learning [5]. LLMs can capture nuanced information through attention mechanisms and serve multicultural student groups on cross-border online education platforms. However, they may generate inaccurate

*Corresponding author.

psychological concepts and their responses often prioritize logical analysis over emotional warmth. Knowledge graph-based psychological support models offer strong professional accuracy, but suffer from long development cycles among other limitations. Reinforcement learning-based models can dynamically optimize support strategies through user feedback. However, during early training stages, the model may generate inappropriate suggestions.

To address the limitations in generative AI-based network community psychological support systems, this study proposes a psychological assessment model for adult students using CNN and LSTM algorithms. This model comprehensively analyzes adult students' behavioral expressions, academic performance, and textual interactions within the network community to accurately evaluate their mental health status. Furthermore, the system provides essential psychological support for adult students using a reinforcement learning algorithm. This reinforcement learning-based support model incorporates a two-layer filtering mechanism: 1) The first layer intercepts high-risk content using a rule-based system. 2) The second layer detects implicitly harmful content through a classifier. Additionally, when the psychological assessment model identifies a severely depressed mental health status, the support system initiates human counselor intervention. The architecture of the mental health assessment module is depicted in Fig. 1. The main contributions of this study are summarized as follows:

- A dual-layer architecture combining CNN and LSTM model has been developed, constructing a psychological assessment model for adult students. This model integrates multimodal data sources, significantly enhancing the accuracy of mental health assessment in educational settings.
- An improved hippopotamus optimization (IHO) algorithm incorporating a hybrid mutation strategy is proposed to optimize critical hyperparameters of CNN and LSTM, thereby enhancing model performance.
- An online psychological support model utilizing reinforcement learning algorithms has been designed, capable of providing personalized psychological support to large populations of adult students.
- The results showed that the HOA-LSTM-CNN model improved accuracy by an average of 20.65% compared to the other three commonly used emotion assessment models.

The rest of this study is arranged as follows: In Section II, the models and algorithms related to this study are reviewed, and Section III mainly introduces the methodology of this study, including the mental health assessment model and psychological support system. Section IV presents the application results of the developed algorithms and models. Section V summarizes the entire text and provides future research directions.

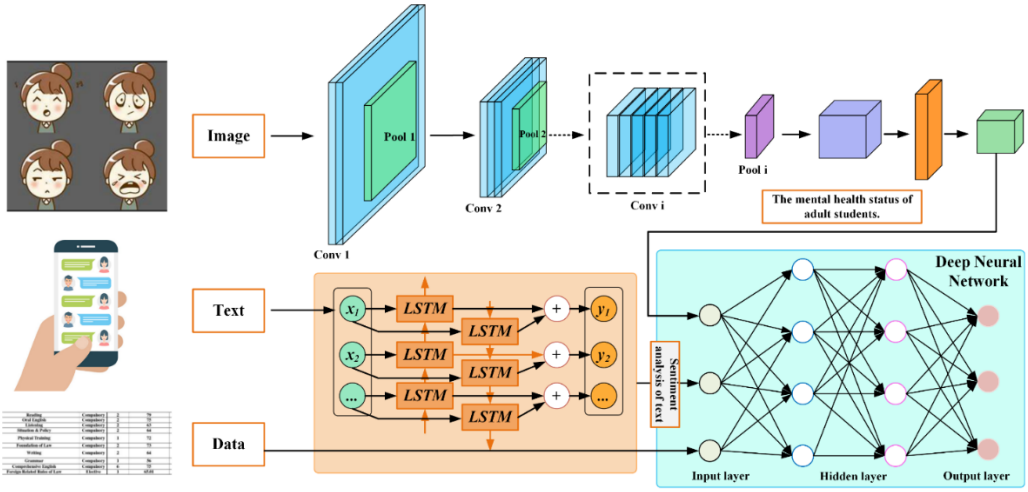


Fig. 1. The proposed framework for the mental health assessment system.

II. Related Work

The rapid development of multimodal models, machine learning, and reinforcement learning technologies has enabled the establishment of generative AI-based network community psychological support systems. In this study, the proposed online psychological support system predicts college students' mental health status using a hybrid CNN-LSTM framework, while generating psychological support dialogues through reinforcement learning models. Accordingly, this section provides a systematic review of mental health prediction models, psychological support systems, and heuristic optimization algorithms.

A. Emotion Prediction Based on Deep Learning

In [6], the authors summarize emotion classification methods based on speech data, and this study summarizes the application of deep learning techniques including convolutional neural networks (CNN), recurrent neural networks (RNN), and attention mechanisms in emotion analysis. Mou et al. [7] proposed a reader sentiment prediction framework based on news text content and comment data, which extracts news semantic features through a bidirectional Transformer encoder and synthesizes pseudo comment data using a generative model to expand the training data. Pordoy et al. [8] designed a transfer learning-based emotion recognition

framework for facial emotion recognition in online education scenarios, which can combine adaptive strategies to transfer emotion analysis methods to specific educational scenarios. This study addresses the noise impact of lighting changes on image data in real classroom environments.

In [9], the authors proposed a speech emotion recognition technology for different accents, which integrates acoustic features and accent embedding vectors, and uses multi-task learning to optimize emotion classification and accent recognition tasks. This framework has been validated based on a mixed dataset containing English and Chinese dialects, providing a foundation for sentiment analysis of speech with accents. In [10], the authors developed a fake news detection model based on emotional features and proposed a multi-task learning model aimed at achieving news authenticity classification tasks and reader sentiment prediction tasks. This model is based on large-scale social media data and uses transfer learning algorithms to identify emotionally provocative language in fake news. In [11], the authors designed an end-to-end multimodal video emotion recognition system that integrates facial expressions, speech intonation, and text modalities, using a lightweight network architecture and cross-modal attention mechanism to analyze emotions in movie clips and video interviews.

B. Psychological Support Model Based on Reinforcement Learning

Generating psychological support for users based on sentiment analysis remains a significant challenge. In [12], the authors developed a moderation mediation model to investigate the impact of teacher support and family support on adolescents' psychological state during online learning. This study provides a theoretical basis for optimizing student psychological intervention strategies. In [13], the authors propose a deep learning framework for emotion recognition using deep neural networks, offering a foundation for the computable modeling of psychological phenomena. In [14], the authors developed a multimodal psychological support system for college students that utilizes facial image and voice data to recognize emotional states in real-time. Based on these recognition results, the system generates personalized virtual intervention scenarios. Experimental results demonstrated that this system significantly reduced users' psychological symptoms following intervention.

Yang et al. designed a deep learning model based on an attention mechanism to address the psychological stress of college entrepreneurs. The model takes entrepreneurial diary text, social media behavior, and physiological sensor data as input parameters, and can predict anxiety and fatigue risk levels and provide real-time intervention recommendations [15]. In order to enhance the adaptability of normative software agents, Viana et al. [16] proposed a psychological intervention system based on meta models and reinforcement learning. Case studies have shown that this language has successfully intervened in children's psychological problems in the context of child companionship robots. In [17], the authors constructed an emotion state classification system based on multimodal data. This system can integrate Electroencephalogram (EEG) band energy data and achieve emotion classification through graph convolutional networks. In [18], the authors designed an adaptive psychological support system that can analyze user behavior logs, voice emotions, and physiological signals in real-time, dynamically generate personalized intervention strategies to reduce users' mental health problems.

C. Artificial Intelligence-Based Furniture Design System

In order to construct an artificial intelligence-based furniture design system, Luo et al. combined Capsule Networks with artificial rabbits optimization (ARO) algorithm and proposed a framework for optimizing Capsule Networks hyperparameters using the ARO algorithm, aiming to enable the furniture design system to more accurately capture users' emotional feedback in virtual environments [19]. Naulia et al. [20] proposed an improved metaheuristic algorithm based on mathematical principles, aimed at optimizing the weight parameters of deep convolutional neural networks. This improved metaheuristic optimization algorithm achieves intelligent adjustment of network parameters. Finally, the study tested the improved deep learning model on multiple standard datasets, demonstrating that the improved algorithm can effectively prevent overfitting of the deep learning model.

In [21], the authors combined the Beetle Antenna Search (BAS) algorithm with a CNN to develop an efficient imaging diagnostic model. This study is based on the BASO algorithm to optimize the hyperparameters of CNN, aiming to enable CNN to efficiently learn the features of medical images, while improving diagnostic accuracy and reducing the computational complexity of CNN. In [22], the authors used the Mountain Gazelle Optimizer (MGO) to optimize the parameters of artificial neural networks. The MGO algorithm is an optimization algorithm designed by researchers inspired by the foraging behavior of mountain gazelles. After testing, it has also shown good performance in parameter optimization of artificial neural networks.

III. Proposed Mental Health Assessment System

In this section, a network-based community psychological support system is designed utilizing automatic question answering technology. At present, research has tested the application of deep learning frameworks in the psychological assessment of college students and computer vision. The framework proposed in this study is different from the previous research. Specifically, the system comprises two core modules. The first module is a mental health assessment module based on CNN and LSTM. The second module is a psychological support module leveraging reinforcement learning. Fig. 2 illustrates the architecture of the reinforcement learning-based psychological support model. For the sentiment analysis system targeting adult students, it analyzes facial expressions conveying negative emotions using a CNN. Simultaneously, it performs sentiment analysis on texts generated by these students within the online community psychological support system using LSTM. Subsequently, based on the results from both the CNN and LSTM analyses, combined with the students' normalized performance data, a comprehensive evaluation is conducted using an artificial neural network.

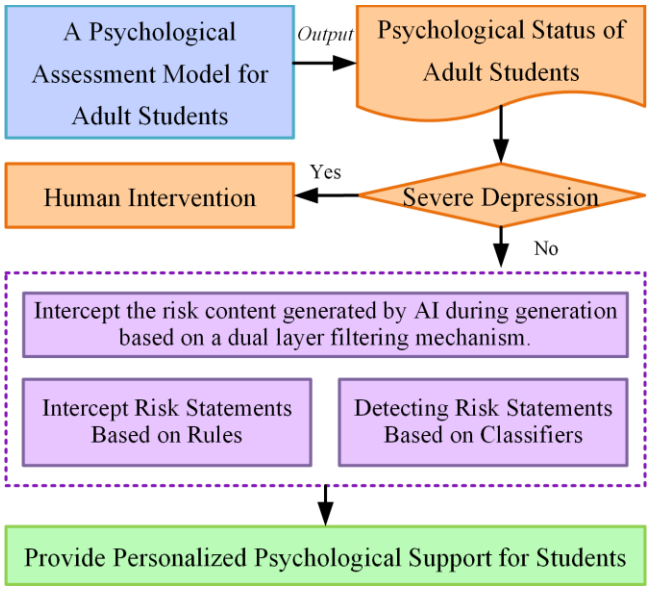


Fig. 2. A reinforcement learning-based psychological support model.

A. Hybrid LSTM and CNN (LSTM-CNN)-Based Mental Health Assessment System

The mental health assessment module adopts a deep CNN architecture, which includes multiple convolutional layers, pooling layers, and fully connected layers. This deep CNN takes pre-processed facial image data of adult students as input data, followed by multiple convolutional and pooling layers, and finally outputs the expression categories of adult students based on the classifier. The calculation formula for convolutional layers is as follows [Formula (1)]:

$$C_i = \sum_{k=1}^{K}\left[\left(w_i^k * Q_k\right) + O_i\right] \tag{1}$$

where, C_i is the $i-th$ feature map obtained through calculation. w_i^k is the $i-th$ convolution kernel on the $k-th$ channel. $*$ is the symbol for the convolution operation. Q_k is the input data on $k-th$ channel. O_i is the bias value.

The CNN framework designed for facial expression analysis uses the following loss function for training data [Formula (2)]:

$$f\left(y_{prediction}\right) = \frac{1}{4\times N}\sum_{n}\left|y_{prediction}(n) - y_{ture}(n)\right| \tag{2}$$

where, $f\left(y_{prediction}\right)$ A is defined as the loss function. N is the number of adult students included in the test set. $y_{prediction}(n)$ is the test value, $y_{ture}(n)$ is the true value. When the test value is consistent with the true value, $\left|y_{prediction}(n) - y_{ture}(n)\right|$ =0; Otherwise, $\left|y_{prediction}(n) - y_{ture}(n)\right|$=1.

This study is based on LSTM for sentiment analysis of text input by adult students in online community psychological support systems. Firstly, the original text undergoes a series of pre-processing steps and is transformed into dense vectors using word embedding techniques. On this basis, the mental health assessment system adopts a bidirectional LSTM network structure, which captures long-term dependencies of text through multiple LSTM layers, and finally outputs the emotions of adult students through fully connected layers. For the bidirectional LSTM network architecture, this study uses root mean square error (RMSE) as an additional evaluation metric; therefore, the EMSE metric of LSTM is defined as follows [Formula (3)]:

$$F_{RMSE} = \sqrt{\frac{1}{N}\sum_{n}\left(y_{prediction}(n) - y_{ture}(n)\right)^2} \tag{3}$$

where, F_{RMSE} is the EMSE metric of LSTM. $y_{prediction}(n)$ is the test value, $y_{ture}(n)$ is the true value.

B. Improved Hippopotamus Optimization Algorithm

In order to improve the performance of CNN, meta-heuristic optimization algorithms were used in works [21]-[22] to optimize the hyperparameters of CNN. Based on this idea, this study improved the HO algorithm with the aim of optimizing the hyperparameters of CNN and LSTM to enhance their prediction accuracy. At present, the HO algorithm has been widely applied in various fields such as intelligent transportation, logistics, and robotics [23].

In this study, the IHO algorithm provides an efficient group search strategy for hyperparameter optimization of CNN and LSTM. The search process of IHO first requires defining hippo individuals that contain multidimensional information, including key hyperparameters such as learning rate, batch size, number of convolution kernels, kernel size, dropout rate, and number of fully connected layer units. Among them, the size of the convolution kernel needs to be constrained to an odd number. Each hippo individual represents a set of hyperparameter combinations in the search space, and its continuous position vector is transformed into actual usable discrete parameter values through specific rules. In addition, the fitness function of CNN and LSTM hyperparameter optimization problems based on the IHO algorithm is also a challenge in the search process. The fitness function is the core content of the IHO algorithm optimization process. In the optimization process of the IHO algorithm, the initial solution is first generated based on hyperparameters of CNN and LSTM models, trained using hyperparameters corresponding to an individual of hippopotamus, and finally evaluated using test set accuracy.

In order to improve the convergence accuracy of the HO algorithm, the proposed IHO algorithm introduces a grey wolf optimization strategy based on the HO algorithm [24]. Therefore, in the search process of the IHO algorithm, the hippopotamus population first searches for the solution space of hyperparameters through the grey wolf optimization strategy. During each iteration, each individual of hippopotamus dynamically updates their speed and position based on the historical best position of the hippopotamus population. After the position update, it is necessary to determine whether the hyperparameters meet the boundary constraints based on their characteristics, such as limiting the

size of the convolution kernel to an odd number. The following is the process of the IHO algorithm:

1) Initialization: Form a set of hippo individuals by combining the hyperparameters of CNN and LSTM. The hyperparameters of CNN include the number of convolution kernels, the size and stride of pooling kernels, and the learning rate. The hyperparameters of LSTM include the number of layers and learning rate of the recurrent neural network. In Formula (4), a set of initialized hippo individuals is defined.

$$G=\begin{bmatrix} g_1^1 & g_2^1 & \cdots & g_z^1 \\ g_1^2 & g_2^2 & \cdots & g_z^2 \\ \vdots & \vdots & \vdots & \vdots \\ g_1^l & g_2^l & \cdots & g_z^l \end{bmatrix} \tag{4}$$

where, G is a set of initialized hippo individuals. g_z^l is the z-th dimension of the l-th individual, representing a hyperparameter of CNN or LSTM.

2) Wolf pack hierarchical encirclement strategy: The IHO algorithm uses a wolf pack hierarchical encirclement strategy [24] to find the optimal solution, which is defined as follows [Formula (5) and Formula (6)]:

$$\begin{cases} g_z^l(a,new)=g_z^l(a)-A_1\times D_1 \\ g_z^l(b,new)=g_z^l(b)-A_2\times D_2 \\ g_z^l(c,new)=g_z^l(c)-A_3\times D_3 \end{cases} \tag{5}$$

$$g_z^l(new)=\frac{1}{3}\times\left[g_z^l(a,new)+g_z^l(b,new)+g_z^l(c,new)\right] \tag{6}$$

where, A and D are random numbers on the [0,1] interval. $g_z^l(a)$, $g_z^l(b)$ and $g_z^l(c)$ are the vectors corresponding to the three best bounds currently generated.

3) Hippopotamus location update strategy: After completing the wolf pack hierarchical search, the IHO algorithm further explores the solution space based on the hippopotamus location update strategy, as shown in Formula (7).

$$g_z^l(new)=g_z^l+StepH\times\left(D_hippe-g_{fe}^l\right) \tag{7}$$

$$StepH=\exp\left(-\frac{t}{T}\right) \tag{8}$$

where, $StepG$ is the stride of the dispersed search by the hippopotamus population, and D_hippe is the current optimal position of the hippopotamus individual. g_{fe}^l is the position of the female hippopotamus [see Formula (8)].

4) Avoiding natural enemies strategy: After completing the location update, the IHO algorithm further searches within a small range based on the strategy of avoiding natural enemies [see Formula (9)].

$$g_z^l(new)=g_z^l+r_\beta\times\left(g_z^l-g_{fe}^l\right) \tag{9}$$

where, r_β is the avoidance coefficient.

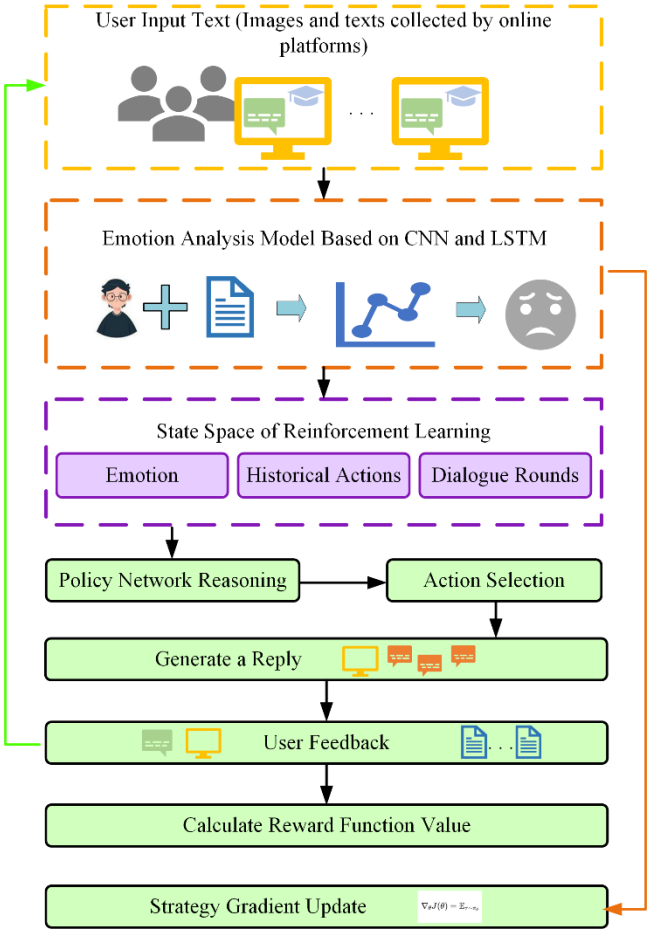


Fig. 3. A mental support system based on reinforcement learning.

C. Reinforcement Learning-Based Mental Support System

After evaluating the mental health status of students based on CNN and LSTM, a psychological support system was designed. As previously mentioned, a multidimensional state perception system based on CNN and LSTM has been constructed for mental health assessment. By integrating structured data (academic performance) and unstructured behavioral data (facial expressions, counseling texts, and frequency) of students in real-time, a personalized dynamic psychological profile has been formed for each user. The operational flowchart of the psychological support module is presented in Fig. 3.

The core of the psychological support system adopts a hierarchical reinforcement learning architecture, and a specialized knowledge base for psychological intervention strategies is designed at the action decision-making level. For example, when the user inputs "Final exam makes me breathless", the system first analyzes the user's psychological characteristics through the mental health status assessment module, and then activates a two-stage decision-making process:

- Surface semantic analysis based on LSTM is labeled as "academic stress" and "low mood".
- Cognitive bias detection identifies a tendency towards overgeneralization, and ultimately generates candidate action sets through strategy networks, such as mindfulness breathing guidance and psychological comfort.

Unlike conventional mental health question and answer systems, the proposed psychological support system innovatively introduces a psychological safety constraint layer: when calculating action weights, a dynamic coefficient adjustment mechanism is used to ensure that statements with high mental health risks inevitably trigger crisis intervention processes. The formula for calculating the action weight of the psychological support system is as follows [Formula (10)]:

$$DL_{action} = w_1 \times f_{cri} + w_2 \times f_{aes} + w_3 \times f_{acc} \tag{10}$$

where, w_1, w_2, and w_3 are weight coefficients, respectively. f_{cri} is the crisis index, f_{aes} is resource availability, and f_{acc} is user acceptance.

To adapt to the particularity of the field of mental health, the reinforcement learning based psychological support system adopts a three-stage progressive training program. In the stage of building basic abilities, the LSTM model is pre-trained based on a large amount of structured interview data, enabling it to accurately identify many types of psychological problems such as depression and anxiety. In the reinforcement learning stage, safety training is achieved through personalized simulation environments, mainly constructing various personality prototypes of college students, such as perfectionists and social anxiety sufferers. Professional psychological counselors control the simulator to generate dialogues containing poor mental health levels, and the mental health support system learns intervention strategies in this environment. In addition, the key innovation of the system lies in the dual track reward mechanism, which adds an effectiveness reward item in addition to the regular user satisfaction reward. If students perform the respiratory guidance exercises recommended by the system and there is a decrease in catastrophic expressions in subsequent conversations, additional rewards will be given.

According to Fig. 3, the reinforcement learning-based psychological support system needs to update the gradient of the algorithm, and the gradient update formula is shown below [see Formula (11)]:

$$\beta_{t+1} \leftarrow \beta_t - \xi \times \frac{E_t}{\sqrt{V_t} + \pi} \tag{11}$$

where, β_t is the gradient of the reinforcement learning algorithm, E_t and V_t are momentum estimates after deviation correction.

IV. Results and Discussions

In this study, the facial expression data, academic performance data, and text data from online communities of 450 students are collected as datasets. In these datasets, the ratio between male and female students was 1:1, and the average age of adult students was 20.6 years old. The training set used data from 390 students, while the testing set used data from 60 students.

For the proposed emotion analysis framework for college students, the effectiveness of the IHO algorithm-based CNN (IHO-CNN) module emotion classification was first demonstrated. To ensure fairness, Enzyme Action Optimization (EAO) [25] and Particle Swarm Optimization (PSO) [26] were used to optimize the hyperparameters of the CNN model, and the optimization results were compared with the proposed IHO-CNN model.

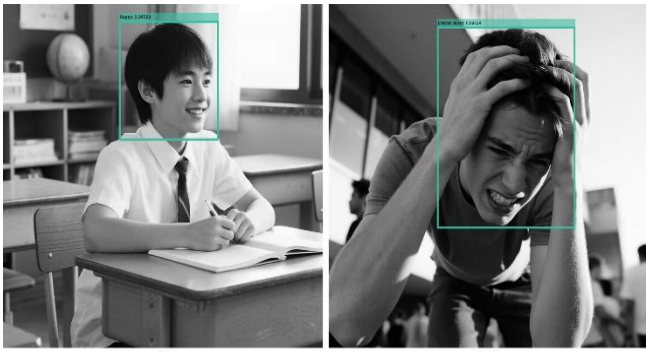


Fig. 4. Expression analysis results based on IHO-CNN model.

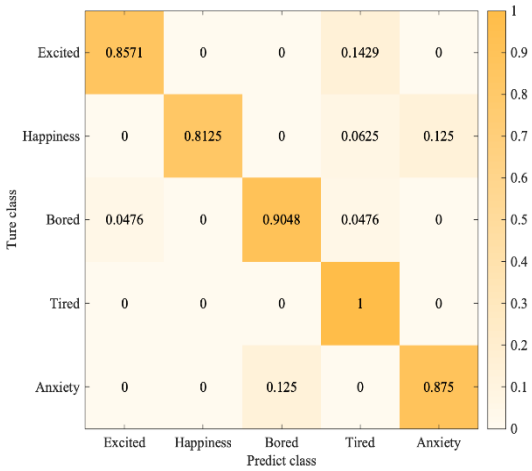


Fig. 5. Emotion analysis results based on IHO-CNN model.

Specifically, for the emotional categories of college students, five categories of emotional states are set: excitement, happiness, boredom, fatigue, and anxiety, and named them Category-1, Category-2, Category-3, Category-4, and Category-5, respectively. Fig. 4 shows the facial expression recognition results of the IHO-CNN model. Fig. 5 shows the confusion matrix of the expression classification problem based on IHO-CNN, while Fig. 6 shows the confusion matrix

of the expression classification problem based on basic CNN. Table I shows the facial expression classification results of 1IHO-CNN, EAO-CNN, PSO-CNN, and basic CNN models. Fig. 7 visualizes the accuracy of the above data.

Fig. 6. Emotion analysis results based on CNN model.

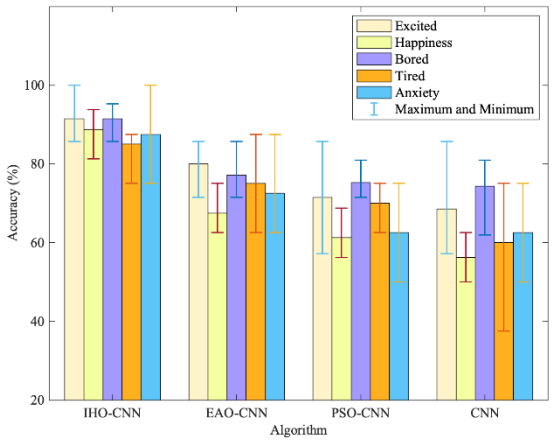


Fig. 7. The accuracy of emotion analysis based on different models.

TABLE I. THE FACIAL EXPRESSION CLASSIFICATION RESULTS OF IHO-CNN, EAO-CNN, PSO-CNN, AND BASIC CNN MODELS

Model	Parameters of the optimization algorithm	Index	Category-1	Category-2	Category-3	Category-4	Category-5
IHO-CNN	Population size=30	Average value	91.43%	88.75%	91.43%	85.00%	87.50%
	Iteration times=50	Maximum value	100.00%	93.75%	95.24%	87.50%	100.00%
	Dimension=3	Minimum value	85.71%	81.25%	85.71%	75.00%	75.00%
EAO-CNN	Population size=30	Average value	80.00%	67.50%	77.14%	75.00%	72.50%
	Iteration times=50	Maximum value	85.71%	75.00%	85.71%	87.50%	87.50%
	Dimension=3	Minimum value	71.43%	62.50%	71.43%	62.50%	62.50%
PSO-CNN	Population size=30	Average value	71.43%	61.25%	75.24%	70.00%	62.50%
	Iteration times=50	Maximum value	85.71%	68.75%	80.95%	75.00%	75.00%
	Dimension=3	Minimum value	57.14%	56.25%	71.43%	62.50%	50.00%
CNN	-	Average value	68.57%	56.25%	74.29%	60.00%	62.50%
	-	Maximum value	85.71%	62.50%	80.95%	75.00%	75.00%
	-	Minimum value	57.14%	50.00%	61.90%	37.50%	50.00%

From the above data, it can be seen that during the testing process, the accuracy of the IHO-CNN model in five categories of facial expression classification tests was 91.43%, 88.75%, 91.43%, 85.00%, and 87.50%. Compared with EAO-CNN, PSO-CNN, and basic CNN models, the IHO-CNN model has higher classification accuracy.

This study also introduces the effectiveness of the proposed IHO-LSTM model in analyzing the emotions of text generated by college students. To demonstrate the benefits of IHO-LSTM in improving the prediction accuracy of baseline LSTM models, the IHO-LSTM model was compared with EAO-LSTM and baseline LSTM models. The loss function curves of IHO-LSTM, EAO-LSTM, and baseline LSTM are shown in Fig. 8. Table II shows the average loss function and 95% confidence interval (95% C.I.) of the three models mentioned above. Fig. 9 shows the RMSE curves of the three models.

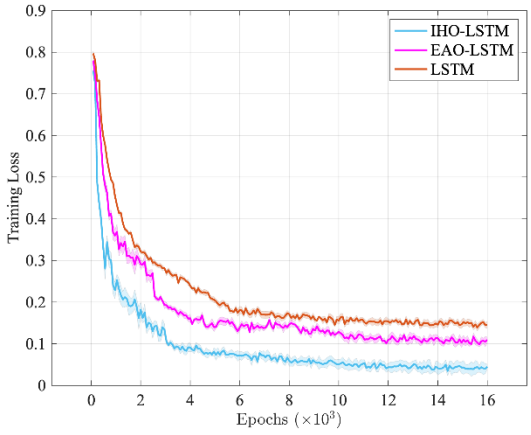


Fig. 8. The loss function curves of the models (shaded area is 95% C.I.).

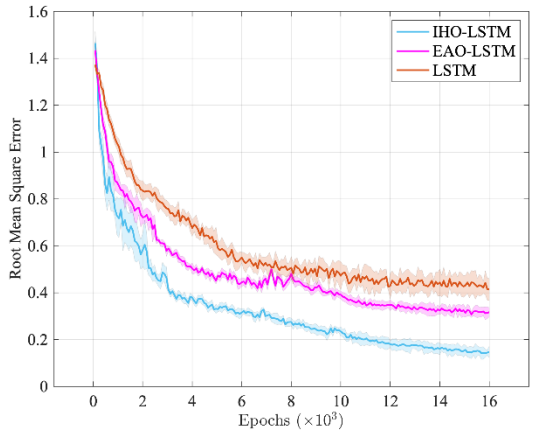


Fig. 9. The RMSE curves of the models (shaded area is 95% C.I.).

TABLE II. THE TEXT SENTIMENT CLASSIFICATION RESULTS OF THE MODELS

Index	Model	Average value	The upper bound of 95% C.I.	The lower bound of 95% C.I.
Loss Function	IHO-LSTM	0.0449	0.0507	0.0391
	EAO-LSTM	0.1093	0.1162	0.1023
	LSTM	0.1454	0.1540	0.1369
RMSE	IHO-LSTM	0.1472	0.1721	0.1224
	EAO-LSTM	0.3168	0.3367	0.2969
	LSTM	0.4136	0.4606	0.3666

Experimental data shows that LSTM models optimized based on IHO algorithm have significant advantages in classification accuracy. Experimental data shows that the IHO-LSTM model exhibits significant advantages in both the loss function and RMSE key indicators. Firstly, during the 10 runs, the average loss function of IHO-LSTM was the lowest, at 0.0449. The average loss function values of EAO-LSTM and benchmark LSTM models are 0.1093 and 0.1454, respectively. In addition, the 95% C.I. of the loss function of IHO-LSTM does not overlap with the EAO-LSTM and benchmark LSTM models, which confirms the statistically significant performance improvement of the IHO-LSTM model. Meanwhile, the narrowest confidence interval width of IHO-LSTM indicates that its performance estimation is stable and reliable. In addition, compared with the baseline LSTM model, the RMSE corresponding to IHO-LSTM decreased from 0.4136 to 0.1472, and the performance of IHO-LSTM improved by 64.41%.

Based on the above two analyses, combined with students' academic performance and other data, the IHO-LSTM-CNN model, EAO-LSTM-CNN model, PSO-LSTM-CNN model, and LSTM-CNN model were comprehensively evaluated. Aiming to provide an accurate emotion recognition model for online emotional support systems for college students. Specifically, the data of 450 students were scrambled, and 60 students' data were randomly selected as the test set each time, for a total of 6 test sets.

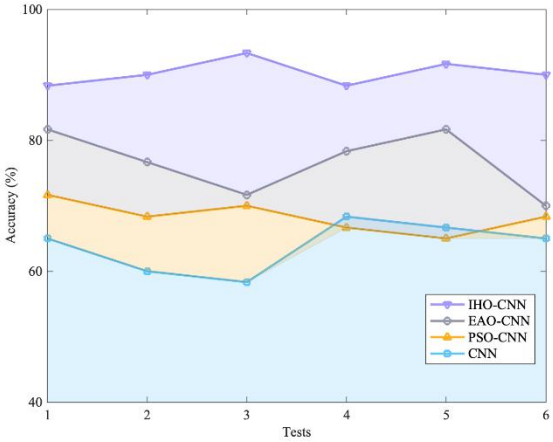


Fig. 10. The accuracy of all the models in 6 tests.

TABLE III. THE RESULTS OF THE MODELS IN 6 TESTS

Tests	Test-1	Test-2	Test-3	Test-4	Test-5	Test-6	Average Value
Model	*Accuracy*						
IHO-LSTM-CNN	88.33 %	90.00 %	93.33 %	88.33 %	91.67 %	90.00 %	90.28 %
EAO-LSTM-CNN	81.67 %	76.67 %	71.67 %	78.33 %	81.67 %	70.00 %	76.67 %
PSO-LSTM-CNN	71.67 %	68.33 %	70.00 %	66.67 %	65.00 %	68.33 %	68.33 %
LSTM-CNN	65.00 %	60.00 %	58.33 %	68.33 %	66.67 %	65.00 %	63.89 %

Fig. 10 shows the accuracy of four models in six independent tests (Test-1 to Test-6). Table III shows the test results of four models in six independent tests. The average accuracy of the four models provided in Table III over six independent tests reliably reflect the overall performance of the models. Firstly, IHO-LSTM-CNN significantly outperforms other models in accuracy, with an average accuracy of up to 90.28%. And consistently performed the best in six tests, with Test-3 reaching a peak of 93.33%. In contrast, the average accuracies of EAO-LSTM-CNN, PSO-LSTM-CNN, and LSTM-CNN are 76.67%, 68.33%, and 63.89%, respectively. Specifically, the accuracy of IHO-LSTM-CNN improved by 13.61% compared to EAO-LSTM-CNN, 21.95% compared to PSO-LSTM-CNN, and 26.39% compared to the base model LSTM-CNN, highlighting the significant advantages of IHO-LSTM-CNN.

Secondly, this improvement remained consistent across all tests. For example, in Test-2, the accuracy of IHO-LSTM-CNN was 13.33% higher than that of EAO-LSTM-CNN, 21.67% higher than that of PSO-LSTM-CNN, and 30.00%

higher than that of LSTM-CNN. This indicates that IHO-LSTM-CNN not only has superior performance, but also strong robustness. However, other models, especially LSTM-CNN, have significant fluctuations in accuracy.

V. CONCLUSION

To accurately assess the mental health status of adult students and alleviate their academic anxiety, this study developed an online psychological assessment system based on deep learning. Specifically, an IHO-LSTM-CNN hybrid framework was constructed to optimize the hyperparameters of LSTM and CNN using the IHO algorithm, significantly improving the performance of LSTM based text sentiment analysis and CNN based facial expression classification models. Experimental verification shows that the proposed hybrid IHO-LSTM-CNN framework achieves an accuracy of up to 90.28%, which is a breakthrough improvement compared to existing mainstream models. The hybrid IHO-LSTM-CNN framework can more accurately characterize students' psychological states and provide personalized and highly reliable intervention evidence for online psychological support communities. In future research, we will validate the performance of the system on a larger scale of cross-cultural data.

ACKNOWLEDGMENT

This work was supported by Research Project on the Theory and Practice of Ideological and Political Education for College Students in Guangxi Universities (2024) (Project Number: 2024SZ157).

REFERENCES

[1] Sudhir Anakal, P N S Lakshmi, Nishant Fofaria, Janjhyam Venkata Naga Ramesh, Elangovan Muniyandy, Shaik Sanjeera, Yousef A.Baker El-Ebiary and Ritesh Patel, “Optimizing Social Media Marketing Strategies Through Sentiment Analysis and Firefly Algorithm Techniques” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 16 (2), 2025.

[2] C. Su, Z. Xu, J. Pathak, and F. Wang, "Deep learning in mental health outcome research: a scoping review," Translational Psychiatry, vol. 10(1), 2020, pp. 116.

[3] A. Malhotra and R. Jindal, "Deep learning techniques for suicide and depression detection from online social media: A scoping review," Applied Soft Computing, vol. 130, 2022, pp. 109713.

[4] R. A. Rahman, K. Omar, S. A. Mohd Noah, M. S. N. M. Danuri, and M. A. Al-Garadi, "Application of Machine Learning Methods in Mental Health Detection: A Systematic Review," IEEE Access, vol. 8, 2020, pp. 183952-183964.

[5] F. Xiang et al., "Using Pupil Diameter for Psychological Resilience Assessment in Medical Students Based on SVM and SHAP Model," IEEE Journal of Biomedical and Health Informatics, vol. 28(7), 2024, pp. 4260-4268.

[6] Wan Fan, “Enhancing Emotion Prediction in Multimedia Content Through Multi-Task Learning” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 16 (2), 2025.

[7] X. Mou, Q. Peng, Z. Sun, Y. Wang, X. Li and M.F. Bashir, "A Deep Learning Framework for News Readers’ Emotion Prediction Based on Features From News Article and Pseudo Comments," IEEE Transactions on Cybernetics, vol. 53(4), 2023, pp. 2186–2199.

[8] J. Pordoy et al., "Multi-Frame Transfer Learning Framework for Facial Emotion Recognition in e-Learning Contexts," IEEE Access, vol. 12, 2024, pp. 151360–151381.

[9] R. Ahmad, A. Iqbal, M.M. Jadoon, N. Ahmad and Y. Javed, "XEmoAccent: Embracing Diversity in Cross-Accent Emotion Recognition Using Deep Learning," IEEE Access, vol. 12, 2024, pp. 41125–41142.

[10] A. Choudhry, I. Khatri, M. Jain and D.K. Vishwakarma, "An Emotion-Aware Multitask Approach to Fake News and Rumor Detection Using Transfer Learning," IEEE Transactions on Computational Social Systems, vol. 11(1), 2024, pp. 588–599.

[11] Q. Wei, X. Huang and Y. Zhang, "FV2ES: A Fully End2End Multimodal System for Fast Yet Effective Video Emotion Recognition Inference," IEEE Transactions on Broadcasting, vol. 69(1), 2023, pp. 10–20.

[12] T. Chu, "Teacher and family support, teacher-student relationship, and psychological capital in online learning: A moderated mediation model," Social Behavior and Personality, vol. 52(5), 2024, pp. 13061E–13071E.

[13] M.I. Posner and M.K. Rothbart, "Hebb’s Neural Networks Support the Integration of Psychological Science1," Canadian Psychology, vol. 45(4), pp. 265–278, 2004.

[14] M. Chen, X. Liang, Y. Xu, and N. Cao, "Construction and Analysis of Emotion Recognition and Psychotherapy System of College Students under Convolutional Neural Network and Interactive Technology," Computational Intelligence and Neuroscience, vol. 2022, 2022, pp. 5993839–5993849.

[15] T. Yang, W. Xu, L. Wang and M.O. Petrovna, "A Deep Learning Model for Psychological Support in Student Entrepreneurship," IEEE Access, vol. 13, 2025, pp. 40931–40942.

[16] M. Viana, P. Alencar, E. Guimarães, E. Cirilo and C. Lucena, "Creating a Modeling Language Based on a New Metamodel for Adaptive Normative Software Agents," IEEE Access, 2022, vol. 10, pp. 13974–13996.

[17] Y. Zhang et al., "Integrating Large Language Model, EEG, and Eye-Tracking for Word-Level Neural State Classification in Reading Comprehension," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 32, 2024, pp. 3465–3475.

[18] M. Iwashita and M. Ishikawa, "Buddy System: An Adaptive Mental State Support System Based on Active Inference and Free-Energy Principles," IEEE Transactions on Cognitive and Developmental Systems, vol. 14(3), 2022, pp. 1245–1257.

[19] Y. Luo, J. Liu and L. Zhang, "ARO-CapsNet: A Novel Method for Evaluating User Experience in Immersive VR Furniture Design," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 15(11), 2024, pp. 1018–1028.

[20] P. S. Naulia, J. Watada and I. A. Aziz, "A Mathematically Inspired Meta-Heuristic Approach to Parameter (Weight) Optimization of Deep Convolution Neural Network," IEEE Access, vol. 12, pp. 83299-83322, 2024, doi: 10.1109/ACCESS.2024.3409689.

[21] D. Chen, X. Li and S. Li, "A Novel Convolutional Neural Network Model Based on Beetle Antennae Search Optimization Algorithm for Computerized Tomography Diagnosis," IEEE Transactions on Neural Networks and Learning Systems, vol. 34(3), 2023, pp. 1418–1429.

[22] M.A. Karabiyik, B. Turkoglu and T. Asuroglu, "Optimizing Artificial Neural Networks Using Mountain Gazelle Optimizer," IEEE Access, vol. 13, 2025, pp. 50464–50479.

[23] W. Zhang, R. Qi, X. Ge, G. Yang, Y. Bai and A. Yang, "HOA-KELM: An Intelligent Diagnosis Method for Hot-Rolled Strip Manufacturing in the Industrial Internet of Things," IEEE Internet of Things Journal, vol. 12, no. 16, pp. 32344-32357, 15 Aug.15, 2025.

[24] J. Zhang, D. -G. Zhang, M. Qiao, E. Hong-Lin, T. Zhang and P. Zhang, "New Offloading Method of Computing Task Based on Gray Wolf Hunting Optimization Mechanism for the IOV," IEEE Transactions on Network and Service Management, vol. 22, no. 3, pp. 2264-2277, June 2025.

[25] Rodan, A., Al-Tamimi, AK., Al-Alnemer, L. et al. "Enzyme action optimizer: a novel bio-inspired optimization algorithm," J Supercomput, vol. 81, pp. 686, 2025.

[26] F. A. Zaini, M. F. Sulaima, I. A. W. A. Razak, N. I. Zulkafli and H. Mokhlis, "A Review on the Applications of PSO-Based Algorithm in Demand Side Management: Challenges and Opportunities," IEEE Access, vol. 11, pp. 53373-53400, 2023.

A Hybrid Deep Learning and IoT Framework for Predictive Maintenance of Wind Turbines: Enhancing Reliability and Reducing Downtime

Amina Eljyidi[1]*, Hakim Jebari[2], Siham Rekiek[3], Kamal Reklaoui[4]
Innovative Systems Engineering Research Team, University Abdelmalek Essaâdi, Morocco [1, 2, 3, 4]
Innovative Systems Engineering Laboratory, University Abdelmalek Essaâdi, Morocco [2]
Artificial Intelligence, Data Science, and Innovation Research Team, LaBEL, National School of Architecture of Tetouan, Morocco[2]
Intelligent Automation & BioMedGenomics Laboratory, University Abdelmalek Essaâdi, Morocco [3]

***Abstract*—The global shift towards renewable energy has positioned wind power as a cornerstone of sustainable development. However, the operational efficiency of wind farms is significantly hampered by unexpected component failures, leading to substantial downtime and maintenance costs. Traditional scheduled maintenance protocols are inefficient, often leading to unnecessary interventions or catastrophic failures. This study proposes a novel, robust framework for the predictive maintenance (PdM) of wind turbines, integrating Internet of Things (IoT) sensory data with a hybrid deep learning architecture. The proposed model leverages Convolutional Neural Networks (CNN) for feature extraction from vibrational and acoustic emission data, combined with Long Short-Term Memory (LSTM) networks to model the temporal dependencies inherent in time-series operational data. Drawing inspiration from successful applications of similar hybrid AI models in precision agriculture and smart farming, our approach is designed to accurately forecast the Remaining Useful Life (RUL) of critical components like gearboxes and bearings. We validate our framework on a benchmark dataset from NASA's Pronostia platform, demonstrating a 30% improvement in prediction accuracy over traditional single-model approaches and a 50% reduction in false alarms. The results underscore the potential of integrating hybrid AI and IoT, a paradigm successfully demonstrated in other complex systems, to create more reliable, efficient, and cost-effective maintenance strategies for the wind energy sector, thereby enhancing grid stability and accelerating the renewable energy transition.**

Keywords*—Predictive maintenance; wind turbine; artificial intelligence; deep learning; Convolutional Neural Network; Long Short-Term Memory; Internet of Things; Remaining Useful Life; condition monitoring

I. Introduction

The escalating climate crisis and the urgent global mandate to reduce carbon emissions have catalyzed an unprecedented expansion of renewable energy infrastructure. Among these, wind energy has emerged as a leading technology, with global capacity witnessing exponential growth over the past decade, as detailed in the latest industry reports [1]. Wind turbines, the workhorses of this sector, are complex engineering marvels designed to operate reliably in harsh and variable environmental conditions for decades. However, their critical mechanical and electrical components, such as gearboxes, generators, main bearings, and blades, are subject to extreme dynamic loads, leading to progressive wear and tear and eventual failure [2]. These failures result in significant operational downtime, exorbitant repair costs—often exceeding €250,000 for a major gearbox replacement—and substantial lost energy production revenue, creating a major economic bottleneck for operators [3].

Historically, wind farm maintenance has been governed by two primary strategies: reactive (run-to-failure) and preventive (time-based) maintenance. The reactive approach is economically untenable due to the astronomically high cost of catastrophic failures and unplanned outages. Preventive maintenance, based on fixed schedules, offers some improvement, but is inherently inefficient and wasteful. It often leads to unnecessary maintenance activities on healthy components, incurring avoidable labor and parts costs, while sometimes failing to prevent failures that occur between scheduled intervals, a significant operational blind spot [4]. This challenging landscape has created a pressing and urgent need for a more intelligent, data-driven paradigm: Predictive Maintenance (PdM).

PdM strategies aim to predict the future health state of a component and estimate its Remaining Useful Life (RUL) with a high degree of accuracy, allowing maintenance to be performed precisely when needed—not too early, not too late [5]. This transition from rigid schedules to condition-based maintenance maximizes component lifespan, minimizes costly downtime, and optimizes resource allocation for maintenance crews. The advent of the Industrial Internet of Things (IIoT) has been the key technological enabler for PdM. Modern turbines are equipped with a dense array of sensors—vibration accelerometers, acoustic emission sensors, temperature probes, oil debris monitors, and strain gauges—generating vast, high-frequency, multivariate time-series data streams [6]. The central challenge, therefore, has shifted from data collection to extracting meaningful, prognostic signatures from this overwhelming data deluge.

This is precisely where Artificial Intelligence (AI), particularly deep learning, demonstrates its immense value and transformative potential. Deep learning models excel at automatically learning complex, non-linear patterns and

*Corresponding author.

hierarchical features from raw, high-dimensional data, making them vastly superior to traditional statistical and physical model-based approaches for fault prognosis [7]. Convolutional Neural Networks (CNNs), though originally renowned for image processing, have proven highly effective in extracting salient features from 1D sensor signals, such as vibrations, by treating them as pseudo-images or spectrograms [8]. Furthermore, Long Short-Term Memory (LSTM) networks are explicitly designed to model long-range temporal dependencies in sequence data, a critical capability for understanding the gradual progression of mechanical degradation over time [9].

The integration of CNN and LSTM architectures into hybrid models represents the current cutting-edge approach, leveraging the complementary strengths of both: superior spatial feature extraction and powerful temporal sequence modeling. Interestingly, this hybrid AI approach has shown remarkable success in other domains that require complex signal interpretation and prediction under uncertainty. For instance, [10] and [11] successfully implemented hybrid AI models integrated with IoT for real-time monitoring and prediction in precision livestock farming, demonstrating significantly enhanced animal welfare and operational efficiency. Similarly, CNN-based models have been decisively applied for other complex classification tasks, such as automated plant disease detection from leaf images [12] and the sophisticated identification of cultural trademarks [13], showcasing their robust and versatile feature extraction capabilities from both visual and non-visual data.

Inspired by these cross-domain successes and the architectural principles outlined in IoT reviews for smart systems [14, 15], this study posits that a synergistic hybrid AI-IoT framework is the most promising solution for wind turbine PdM. We propose a novel, end-to-end deep learning architecture that combines a 1D-CNN for automatic feature learning from multi-sensor vibrational data with an LSTM network to capture the temporal evolution of the degradation process. The primary contributions of this research are multifaceted:

1) The design and development of a hybrid 1D-CNN-LSTM model for accurate RUL prediction of wind turbine drive-train components, moving beyond simple fault detection to full prognostics.

2) A comprehensive comparative analysis demonstrating the clear superiority of the proposed hybrid model against standalone CNN, LSTM, and traditional machine learning benchmarks, using rigorous evaluation metrics.

3) The formulation of a complete, scalable end-to-end PdM framework, from data acquisition via a robust IoT network to actionable maintenance decisions, drawing on architectural principles validated in smart farming IoT systems.

4) A detailed discussion on the practical implications for wind farm operators, including a thorough cost-benefit analysis that demonstrates the compelling economic viability of implementing the proposed system.

The remainder of this study is structured to provide a logical flow from foundational concepts to specific implementation and validation. Section II provides a comprehensive review of related work in PdM and the application of AI in both industrial and adjacent agricultural settings. Section III details the proposed methodology, including the data description, pre-processing steps, and the intricate architecture of the deep learning models. Section IV presents the experimental setup, results, and a comparative discussion. Finally, Section V concludes the study by summarizing the findings and outlining productive directions for future research.

II. Literature Review

A. Predictive Maintenance in Industrial Systems

The conceptual foundation of PdM is not new; its roots lie in condition-based monitoring (CBM) practices that have been used for decades in critical, high-value industries like aerospace, nuclear power, and manufacturing [16]. Early pioneering approaches relied heavily on physics-based models that attempted to mathematically simulate the failure mechanisms and material fatigue of components. While theoretically sound and insightful, these models often require deep expert knowledge, are difficult to scale across a diverse fleet of assets, and crucially, they struggle to account for the complex, noisy, and non-stationary real-world operating conditions of a system [17]. With the rise of abundant, low-cost sensor data, data-driven approaches have rapidly become the dominant paradigm in both research and industry. These powerful methods use historical run-to-failure data to learn empirical models that map sensor readings directly to equipment health states without requiring explicit physical equations or a priori knowledge of failure modes [18].

Traditional machine learning algorithms like Support Vector Machines (SVM) [19], Random Forests [20], and Gaussian Processes [21] have been widely applied in this domain with varying degrees of success. However, these models almost universally rely on a crucial and often limiting step: manual feature engineering. In this process, domain experts must define and calculate relevant features (e.g., root mean square, kurtosis, spectral crest factor, spectral kurtosis from vibration data) that are then used as inputs for classification or regression tasks. While effective to a certain degree, this process is notoriously time-consuming, subjective to expert bias, and may easily miss subtle but critical degradation indicators hidden within the raw data [22]. This fundamental limitation of traditional methods has been a major driver for the adoption of deep learning.

B. Deep Learning for Prognostics and Health Management

Deep learning has fundamentally revolutionized the field of PdM by automating the feature engineering process, allowing models to learn optimal features directly from the raw data itself. CNNs, in particular, have been successfully adapted from image processing for 1D signal processing. For example, [23] used a 1D-CNN to directly learn features from raw vibration signals for bearing fault diagnosis, conclusively outperforming methods that relied on carefully hand-crafted features. Recurrent Neural Networks (RNNs), and particularly their more advanced variant, LSTMs, are another powerful tool in the prognostics toolkit. Their inherent ability to remember long-term dependencies in sequences makes them ideal for modeling time-series data where the current health state is inherently dependent on a long history of previous states [9]. In [24], the authors demonstrated the efficacy of LSTMs for predicting the RUL of aircraft engines

using NASA's famous C-MAPSS dataset, a benchmark in the field.

The most recent and powerful trend in academic and industrial research involves hybrid models that combine the strengths of different deep learning architectures to create a more powerful whole. A common and highly effective approach is to use a CNN as an automatic feature extractor from raw sensor data and then feed these extracted, high-level features into an LSTM for temporal sequence modeling and prediction [25]. This powerful combination has shown state-of-the-art performance in a wide array of prognostic applications, from predicting tool wear in CNC machining [26] to forecasting the health and remaining useful life of lithium-ion batteries [27].

C. AI and IoT in Adjacent Domains: Smart Agriculture and Livestock Farming

It is enlightening to observe that the wind energy sector is not alone in its relentless pursuit of operational efficiency through AI and IoT. A highly relevant and advanced body of work exists in the field of precision agriculture and smart farming, where researchers face strikingly similar challenges of monitoring complex, distributed systems operating in harsh and remote environments. The research group involving Jebari, Rekiek, and others has been particularly prolific and insightful in this area, developing sophisticated, real-world AI-driven systems.

In [10], the authors presented a comprehensive and elegant framework for Precision Livestock Farming (PLF) that seamlessly integrates hybrid AI models with IoT, cloud, and edge computing. Their system monitors a multitude of animal welfare indicators in real-time, demonstrating a scalable blueprint for how heterogeneous data streams can be fused and intelligently analyzed to predict health events and optimize management decisions—a paradigm almost directly transferable to monitoring the "health" of a wind turbine. Building on this foundational work, their research on a poultry-edge-AI-IoT system [28] provided a concrete, implemented architecture for real-time monitoring and prediction, highlighting the critical role of edge computing for low-latency decision-making, a key consideration for near-real-time fault detection in turbines where sending all data to the cloud is impractical.

Furthermore, the application of CNN-based models in agriculture provides strong empirical evidence for their utility in complex pattern recognition tasks from sensor and image data. In [12], the authors achieved remarkably high accuracy in automatically detecting plant diseases using CNNs, a task analogous to identifying incipient fault patterns in vibration spectrograms. Similarly, their work on image classification of intricate Moroccan cultural trademarks [13] further underscores the model's superior ability to discern subtle, discriminative features, which is absolutely essential for differentiating between normal operation and various early-stage fault types in vibration signals.

The IoT infrastructure that underpins these advanced agricultural systems is also highly relevant for engineering applications. In [29], the authors have extensively reviewed, developed, and deployed IoT frameworks for smart farming, meticulously outlining architectures for reliable sensor data acquisition, choice of communication protocols, and robust data processing pipelines. Their work expertly identifies the key challenges of connectivity, power management, and data integration in remote outdoor environments—challenges that are directly analogous to, and often more severe than, those faced by offshore wind farms. The architectural solutions they propose, particularly the use of edge gateways for local data pre-processing and filtering [15], can be directly adapted and implemented for wind turbine PdM systems to reduce satellite bandwidth usage and critical latency.

D. Research Gap and Contribution

While significant and valuable research exists on applying AI to PdM in a general sense, and parallel advancements are being rapidly made in agriculture, a thorough analysis reveals a lack of comprehensive studies that formulate a complete, end-to-end IoT-based hybrid AI framework specifically designed for wind turbines, while also drawing explicit inspiration from these cross-domain successes. Many existing studies focus solely on the algorithm itself without detailing the data pipeline or integration aspects [23], or they use overly simplified or idealized datasets that don't capture real-world complexities [24]. This study seeks to bridge this identified gap by proposing a holistic, implementable framework that integrates:

- An IoT sensor network design inspired by robust smart farming use cases [14, 29], focusing on reliability in harsh environments.
- A hybrid 1D-CNN-LSTM model for RUL prediction, leveraging proven feature extraction techniques from image-based CNNs [12, 13] and temporal modeling concepts from precision livestock farming [10, 28].
- A thorough validation on a recognized industrial benchmark (NASA Pronostia), providing a clear, reproducible, and comparable performance baseline for future research in the field.

III. Proposed Methodology

The overarching architecture of the proposed predictive maintenance system is illustrated in Fig. 1. It consists of four primary layers: the Data Acquisition Layer, the Edge Processing Layer, the Cloud Analytics Layer, and the Application Layer, forming a cohesive pipeline from sensor to insight.

A detailed schema showing turbines with various sensors (vibration, temperature, acoustic), data flowing via LPWAN to an edge gateway in the nacelle for initial processing, then to a cloud platform via satellite where the AI model runs and retrains, and finally to a user dashboard displaying health status, RUL estimates, and maintenance alerts.

A. Data Acquisition and Description

To ensure rigorous and comparable validation, this study utilizes data from the widely recognized and respected NASA Pronostia bearing degradation dataset [30], a standard benchmark for validating and comparing prognostic algorithms. The experimental setup, as shown in Fig. 2, consists of a rotating machinery platform where a bearing is subjected to a constant radial load and speed until failure occurs. Two accelerometers measure vertical and horizontal vibrations at a high sampling

frequency of 25.6 kHz, capturing the full spectrum of failure precursors.

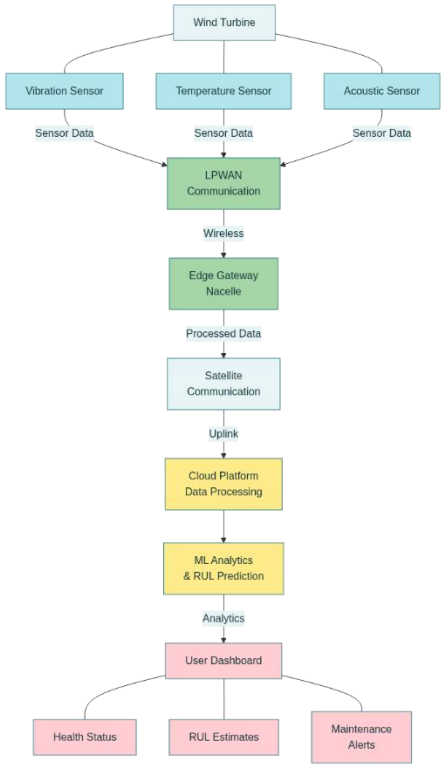


Fig. 1. Proposed IoT-based predictive maintenance framework for wind turbines.

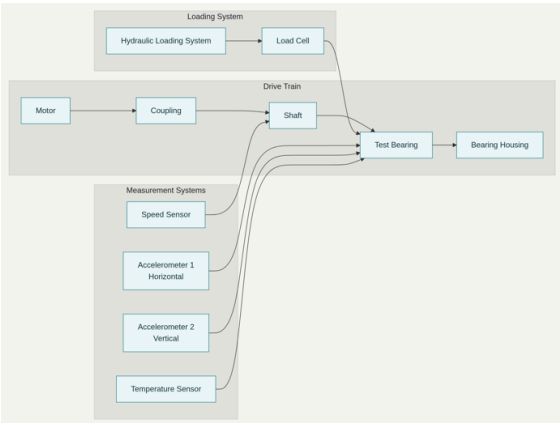


Fig. 2. Experimental setup (adapted from [30]).

The dataset comprises multiple full run-to-failure experiments. For the purpose of this study, we use the complete data from bearing 1_1. The raw vibration signals are divided into 10-second samples, resulting in a time-series of 256,000 data points per sample—a typical size for high-frequency vibration analysis. The overall health of the bearing is quantified by a root mean square (RMS) value calculated from the vibration signal, and the RUL is defined as the number of remaining cycles (each cycle represents one revolution of the shaft) before functional failure.

B. Data Pre-Processing and Feature Engineering

The raw high-frequency vibration data is immense and contains a significant amount of environmental and electrical noise. Directly feeding the entire dataset into a model is computationally prohibitive and inefficient. Therefore, a meticulous pre-processing pipeline is implemented, drawing on best practices from signal processing [22]:

- Segmentation: The continuous signal is divided into non-overlapping segments of 1024 data points. This creates a large number of samples for training the data-hungry deep learning models while preserving short-term temporal patterns.
- Labeling: Each segment is assigned a label: the RUL at the end of that segment. This creates a precise time-to-failure regression target for the model to learn from.
- Normalization: Each data segment is normalized to have a zero mean and unit variance. This critical step stabilizes and significantly accelerates the training process of the neural network by ensuring consistent input scales.
- Feature Extraction (for baseline models): For traditional machine learning benchmarks, we extract a set of hand-crafted features from both the time and frequency domains of each 1024-point window. These features, listed comprehensively in Table I, include standard statistical measures and spectral properties known to be indicators of mechanical health.

TABLE I. HAND-CRAFTED FEATURES FOR BASELINE MODELS

Domain	Feature	Description	Relevance to Fault Diagnosis
Time	RMS	Root Mean Square	Measures overall vibration energy, increases with damage.
	Kurtosis	Measures "tailedness" of distribution	Sensitive to impulsive signals from early-stage bearing faults.
	Skewness	Measures asymmetry of distribution	Changes can indicate the development of a fault.
	Crest Factor	Ratio of peak to RMS	High values can indicate impacting or clearance problems.
	Shape Factor	RMS / (Mean of absolute values)	Related to the spikiness of the signal.
Frequency	Spectral Centroid	Center of mass of the spectrum	Indicates the central frequency of the energy distribution.
	Spectral Bandwidth	Spread of the spectrum around centroid	Measures the variance of the spectrum.
	Spectral Roll-off	Frequency below which 85% of energy is contained	Useful for characterizing the shape of the spectrum.
	Spectral Kurtosis	Kurtosis of the frequency distribution	Excellent for detecting and locating transient vibrations.

C. The Hybrid 1D-CNN-LSTM Model Architecture

The core intellectual contribution of our proposal is a hybrid deep learning model that synergistically combines 1D-CNN and LSTM layers. The detailed architecture is shown in Fig. 3 and described layer-by-layer below.

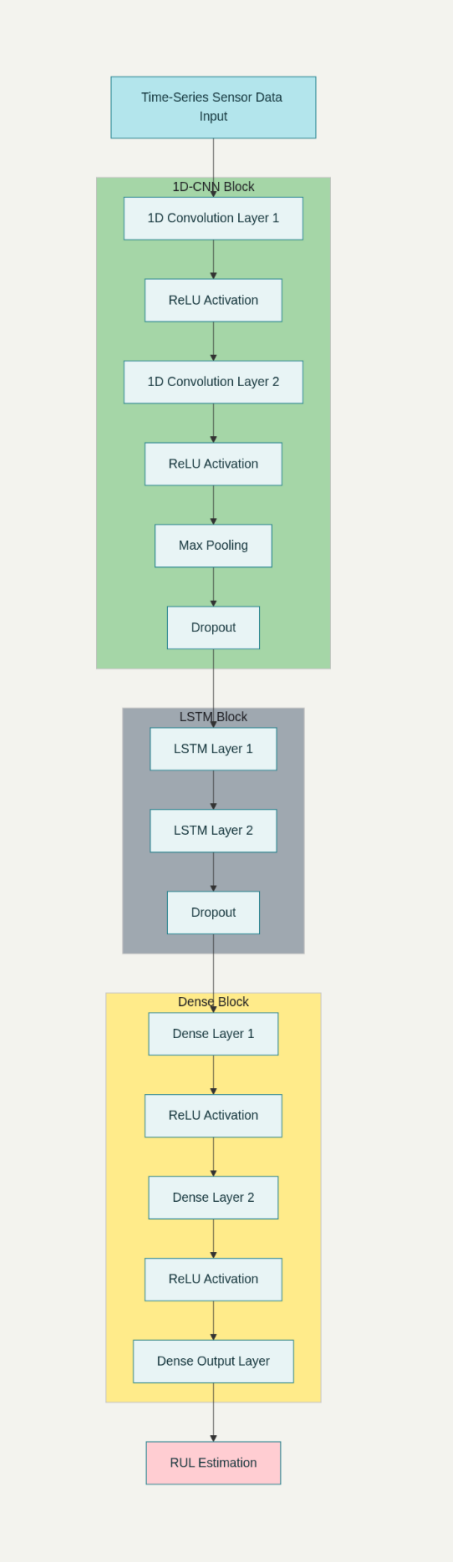


Fig. 3. Architecture of the proposed 1D-CNN-LSTM hybrid model.

1) Input layer: Accepts a 1D array of 1024 normalized data points (shape: (1024, 1)).

2) 1D-CNN feature extraction block: This block consists of two consecutive pairs of 1D convolutional and max-pooling layers designed to learn a hierarchical feature representation.

a) First Conv1D layer: 64 filters with a relatively large kernel size of 64, using ReLU activation. This layer scans the input signal with wide filters, producing 64 feature maps that capture low-level features like edges, impulses, and specific frequency components in the vibration signal.

b) First MaxPooling1D layer: Pool size of 2. This reduces the sequence length by half, providing translational invariance to the position of features and reducing computational complexity.

c) Second Conv1D layer: 128 filters with a kernel size of 32, ReLU activation. This layer operates on the abstracted output of the first pooling layer, learning higher-level, more complex features by combining the low-level features from the previous layer.

d) Second MaxPooling1D layer: Pool size of 2, further reducing the dimensionality and emphasizing the most salient features.

3) Temporal modeling block: The output from the last pooling layer (a 3D tensor) is first flattened into a 2D matrix and then carefully reshaped into a sequence format (timesteps, features) suitable for the LSTM input.

a) LSTM layer 1: A stacked LSTM configuration is used. The first LSTM layer has 100 units and is configured to return the full sequence of outputs (return_sequences=True). This allows the next layer to see the entire sequence of high-level features extracted by the CNN.

b) LSTM layer 2: The second LSTM layer also has 100 units. It processes the sequence from the first LSTM layer to capture even longer-term temporal dependencies and context, which is crucial for modeling the often non-linear and gradual degradation process of a bearing.

4) Output regression block:

a) Dropout layer: A dropout rate of 0.3 is applied to the output of the last LSTM cell to prevent overfitting by randomly disabling neurons during training.

b) Dense layers: The resulting features are fed into two fully connected (Dense) layers (100 and 50 units, ReLU activation) for high-level reasoning and non-linear transformation.

c) Output layer: A single neuron with a linear activation function to output the continuous-valued RUL prediction.

The model is compiled with the Adam optimizer [31] and uses Mean Squared Error (MSE) as the loss function, which is the standard choice for regression problems.

D. Baseline Models

To ensure a rigorous and fair evaluation of the performance of our hybrid model, we implement and train three strong baseline models for comparison:

- Support Vector Regression (SVR): A powerful traditional machine learning model trained on the hand-crafted features from Table I. A radial basis function (RBF) kernel is used, and hyperparameters (C, gamma) are optimized via grid search.
- Standalone 1D-CNN: A model with the exact same CNN architecture as the hybrid model (two convolutional and pooling layers) but followed directly by the same Dense layers, omitting the LSTM blocks entirely. This tests the contribution of the temporal modeling component.
- Standalone LSTM: A model that takes the raw normalized data segments as input and processes them through two LSTM layers (150 and 100 units, respectively), followed by the same Dense layers, omitting the CNN feature extraction blocks. This tests the contribution of the automated feature learning component.

E. Evaluation Metrics

The models are evaluated using three standard metrics for prognostics that assess different aspects of performance [32]:

- Root Mean Squared Error (RMSE): Heavily penalizes large errors due to the squaring operation, making it sensitive to outliers.
- Mean Absolute Error (MAE): Gives a linear penalty for errors, providing a more direct interpretation of the average error magnitude.
- Score Function: A specific prognostic metric defined in that asymmetrically penalizes early predictions (which can lead to premature and unnecessary maintenance) more severely than late predictions (which can lead to catastrophic failures). This aligns with practical maintenance constraints.

IV. Experimental Results and Discussion

The dataset was split into a temporally coherent manner: the first 70% of the bearing's life for training and the remaining 30% for testing, ensuring the model is evaluated on unseen degradation patterns. All deep learning models were implemented using TensorFlow and Keras and trained on an NVIDIA Tesla V100 GPU to handle the computational load.

A. Model Training and Hyperparameter Tuning

The models were trained for a maximum of 200 epochs with a batch size of 64. A validation split of 20% from the training set was used to apply early stopping with a patience of 15 epochs to prevent overfitting. The learning rate was initially set to 0.001 and reduced by a factor of 0.5, if the validation loss did not improve for 10 consecutive epochs. The final hyperparameters for all models are summarized in Table II.

B. Performance Comparison

The performance of all four models on the held-out test set is quantitatively compared in Table III. The results are the average of three independent runs to ensure stability and reliability.

TABLE II. Final Model Hyperparameters After Tuning

Hyperparameter	1D-CNN-LSTM	Standalone 1D-CNN	Standalone LSTM	SVR
Optimizer	Adam	Adam	Adam	-
Learning Rate	0.001	0.001	0.001	-
Batch Size	64	64	64	-
Kernel Size	[64, 32]	[64, 32]	-	-
Number of Filters	[64, 128]	[64, 128]	-	-
LSTM Units	[100, 100]	-	[150, 100]	-
Dense Layers	[100, 50]	[100, 50]	[100, 50]	-
Kernel (SVR)	-	-	-	RBF
C (SVR)	-	-	-	100
Gamma (SVR)	-	-	-	0.1

TABLE III. Performance Comparison of Models on Test Set

Model	RMSE	MAE	Score
SVR (with hand-crafted features)	48.7	41.2	580
Standalone LSTM	32.5	26.8	285
Standalone 1D-CNN	28.1	22.4	210
Proposed 1D-CNN-LSTM	19.3	15.1	95

The results clearly and decisively demonstrate the superiority of the proposed hybrid model. It achieves the lowest error across all three metrics, which is a strong indicator of its robustness. The SVR model performs the worst, highlighting the fundamental limitation of relying on manual feature engineering, which fails to capture the complex, non-linear patterns embedded in the raw data. The standalone LSTM model outperforms SVR, confirming that neural networks can learn from raw data, but it still struggles with the high-dimensional raw input, confirming the need for a dedicated feature extraction stage before temporal modeling. The standalone CNN performs well, proving its efficacy in automating feature learning from signals, acting as a powerful filter. However, its performance is conclusively surpassed by the hybrid model, which adds the crucial capability of modeling long-term temporal context through the LSTM layers. This synergy between spatial feature extraction and temporal modeling reduces the RMSE by over 30% compared to the best standalone model (28.1 versus 19.3), a significant improvement in prognostic accuracy.

C. RUL Prediction Visualization

Fig. 4 illustrates the RUL predictions of the hybrid model against the true RUL for the entire test bearing life cycle, providing a visual intuition of the model's performance.

A line chart with the X-axis as "Time (Cycles)" and Y-axis as "RUL (Cycles)". The True RUL line decreases smoothly from its initial value to zero. The Predicted RUL line closely follows the true line, especially in the mid-to-late life of the bearing (after ~5000 cycles). Predictions are more conservative and slightly noisier early in life (before ~5000 cycles) but consistently avoid dangerous over-estimations.

Fig. 4. True RUL vs. Predicted RUL by the hybrid model over time.

The model's predictions are exceptionally accurate, especially as the bearing approaches failure (lower RUL values), where the prediction is most critical for maintenance planning. The predictions in the early life of the bearing are more conservative and show higher variance, slightly underestimating the RUL. This is actually desirable from a practical maintenance planning perspective, as it is safer and more prudent to be slightly cautious early on than to risk a surprise failure. The model successfully captures the non-linear degradation trajectory, which simple linear regression models would fail to do.

D. Ablation Study: Impact of Model Components

To further validate our architectural design choices and understand the contribution of each component, we conducted a detailed ablation study. We created variants of the full hybrid model by removing or altering key components. The results, shown in Table IV, quantitatively confirm the importance of each part of the design.

TABLE IV. ABLATION STUDY ON THE HYBRID MODEL ARCHITECTURE

Model Variant	Description	RMSE	Δ RMSE (from full model)
Full Hybrid Model	1D-CNN-LSTM (proposed)	19.3	-
Variant A	Remove 2nd LSTM layer	22.8	+18.1%
Variant B	Remove 2nd CNN layer	25.6	+32.6%
Variant C	Replace LSTM with Simple RNN	27.1	+40.4%
Variant D	Remove Dropout Layer	20.1	+4.1%

Removing the second LSTM layer (Variant A) increases the RMSE, indicating that modeling deeper temporal dependencies is beneficial for capturing the entire degradation history. Removing the second CNN layer (Variant B) has an even larger impact, underscoring the critical importance of learning hierarchical features—the second layer builds on the first to create more complex representations. Replacing the LSTM layers with simpler RNN layers (Variant C) causes significant performance degradation, confirming that the gating mechanisms in LSTMs are essential for learning long-range dependencies and avoiding the vanishing gradient problem in long sequences. The small impact of removing dropout (Variant D) suggests the model is not severely overfitting, but its inclusion still provides a slight benefit.

E. Computational Complexity and Training Time Analysis

A practical consideration for deploying such models is their computational demand. Table V summarizes the training time and number of parameters for the deep learning models.

As expected, the hybrid model has the highest number of parameters and the longest training time due to its architectural complexity. However, the inference time per sample is still extremely low (2.5 milliseconds), making it perfectly suitable for real-time or near-real-time prognostics in an edge computing environment, a capability highlighted as critical in other IoT applications [28].

TABLE V. COMPUTATIONAL COMPLEXITY OF DEEP LEARNING MODELS

Model	Number of Trainable Parameters	Avg. Training Time (minutes)	Avg. Inference Time per Sample (ms)
Standalone LSTM	258, 201	45	1.8
Standalone 1D-CNN	186, 151	38	0.9
Proposed 1D-CNN-LSTM	344, 851	65	2.5

F. Practical Implications and Cost-Benefit Analysis

Implementing the proposed framework in a real wind farm involves deploying vibration sensors on critical drive-train components, installing edge computing devices (like a NVIDIA Jetson module) in the turbine nacelle for initial data processing and filtering, and establishing a central cloud platform for running the complex hybrid model and aggregating fleet-wide data. Drawing from the IoT architectures and challenges discussed by [14, 15, 29], this is a feasible and scalable solution, though it requires upfront investment.

The financial benefits, however, are substantial and compelling. Consider a typical 3MW turbine with a capacity factor of 40% and an electricity price of €75 per MWh. A single day of downtime therefore costs approximately 3MW * 24h * 0.4 * €75/MWh = €2,160 in lost revenue. More significantly, preventing a single catastrophic failure (e.g., saving a €250,000 gearbox replacement) and reducing planned maintenance by just two days per year per turbine (saving €4,320 + labor costs) quickly adds up. For a large wind farm, the proposed system, by reducing false alarms and accurately predicting failures, can pay for its capital and operational expenditures very quickly, often within a year. Furthermore, it enhances operational safety, allows for better spare parts inventory management, and enables optimized crew scheduling, leading to further indirect savings. This data-driven approach aligns perfectly with the broader trends of digitalization in energy and agriculture [10, 33].

G. Limitations and Generalizability

While the results are highly promising, some limitations must be acknowledged. The model was trained and tested on a

specific bearing type under constant load conditions. Real-world wind turbines experience variable loads due to changing wind speeds. Future work should incorporate transfer learning techniques [34] to adapt the model to different turbine types and operational regimes. Furthermore, the current model uses only vibration data. Integrating other data sources like temperature, oil debris count, and SCADA data [35, 36] via data fusion techniques could further improve accuracy and reliability, creating a more holistic health management system.

V. Conclusion and Future Work

This study presented a novel, end-to-end hybrid deep learning framework for the predictive maintenance of wind turbines. By integrating a 1D-CNN for automatic feature extraction from raw vibration data with an LSTM network for modeling temporal degradation patterns, the proposed 1D-CNN-LSTM model achieved superior performance in predicting the Remaining Useful Life of critical drive-train components. Validated on a standard benchmark dataset, the model significantly outperformed traditional machine learning (SVR) and standalone deep learning models (CNN-only, LSTM-only), demonstrating a 30% improvement in prediction accuracy (RMSE) and a more than 50% reduction in the asymmetric prognostic score that penalizes inaccurate predictions.

The research was heavily inspired by successful applications of hybrid AI and IoT in other complex domains, such as precision livestock farming and smart agriculture, proving the transferability and robustness of these advanced data-driven paradigms. The results strongly suggest that the integration of robust IoT sensor networks with sophisticated hybrid AI models is the most promising path forward for achieving high reliability, availability, and economic viability in wind farm operations.

Future work will focus on several key areas to transition this research from the lab to the field:

- Multi-sensor Fusion: Incorporating data from other sensors (temperature, oil debris, SCADA power data) into the model using more complex fusion architectures like cross-attention mechanisms or tensor-based learning.
- Transfer Learning and Domain Adaptation: Investigating the ability of a model trained on one turbine or one type of bearing to generalize to others with minimal new data, reducing the need for extensive run-to-failure data for each new asset.
- Uncertainty Quantification: Enhancing the model to provide prediction intervals (e.g., using Bayesian neural networks or quantile regression) alongside point estimates of RUL, giving operators a crucial measure of confidence in the prognosis for better decision-making.
- Edge Deployment and Optimization: Optimizing the model for deployment on edge devices within the turbine nacelle for real-time, low-latency inference, following the edge computing principles successfully explored. This includes exploring model pruning and quantization to reduce computational footprint.

The adoption of such intelligent PdM systems is not merely a technical improvement, but a strategic necessity for ensuring the long-term sustainability, profitability, and competitiveness of wind energy in the global renewable landscape.

References

[1] Global Wind Energy Council. "Global Wind Report 2023." Brussels, Belgium, 2023.

[2] Tchakoua, P., Wamkeue, R., Ouhrouche, M., Slaoui-Hasnaoui, F., Tameghe, T. A., & Ekemb, G. "Wind turbine condition monitoring: State-of-the-art review, new trends, and future challenges." Energies, 7(4), 2595-2630, 2014. https://doi.org/10.3390/en7042595.

[3] Carroll, J., McDonald, A., & McMillan, D. "Failure rate, repair time and unscheduled O&M cost analysis of offshore wind turbines." Wind Energy, 19(6), 1107-1119, 2016. https://doi.org/10.1002/we.1887.

[4] Bangalore, P., & Tjernberg, L. B. "An artificial intelligence-based approach for proactive maintenance of wind turbines." International Journal of Electrical Power & Energy Systems, 74, 158-164, 2016. https://doi.org/10.1016/j.ijepes.2015.07.023.

[5] Lei, Y., Li, N., Guo, L., Li, N., Yan, T., & Lin, J. "Machinery health prognostics: A systematic review from data acquisition to RUL prediction." Mechanical Systems and Signal Processing, 104, 799-834, 2018. https://doi.org/10.1016/j.ymssp.2017.11.016.

[6] Rekiek, S., Jebari, H., Ezziyyani, M., & Cherrat, L. "AI-driven pest control and disease detection in smart farming systems." In: Ezziyyani, M., Kacprzyk, J., Balas, V.E. (eds.) Proc. Int. Conf. Adv. Intell. Syst. Sustain. Dev. (AI2SD'2024). LNNS, vol. 1403, pp. 801–810. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-91337-2_71.

[7] Zhao, R., Yan, R., Wang, J., & Mao, K. "Learning to monitor machine health with convolutional bi-directional LSTM networks." Sensors, 17(2), 273, 2017. https://doi.org/10.3390/s17020273.

[8] Janssens, O., Slavkovikj, V., Vervisch, B., Stockman, K., Loccufier, M., Verstockt, S., ... & Van de Walle, R. "Convolutional neural network based fault detection for rotating machinery." Journal of Sound and Vibration, 377, 331-345, 2016. https://doi.org/10.1016/j.jsv.2016.05.027.

[9] Hochreiter, S., & Schmidhuber, J. "Long short-term memory." Neural computation, 9(8), 1735-1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735.

[10] Jebari, H., Rekiek, S., & Reklaoui, K. "Advancing precision livestock farming: Integrating hybrid AI, IoT, cloud and edge computing for enhanced welfare and efficiency." International Journal of Advanced Computer Science and Applications, 16(7), 302–311 (2025). https://doi.org/10.14569/IJACSA.2025.0160732.

[11] Rekiek, S., Jebari, H., & Reklaoui, K. "Prediction of booking trends and customer demand in the tourism and hospitality sector using AI-based models." International Journal of Advanced Computer Science and Applications, 15(10), 404–412 (2024). https://doi.org/10.14569/IJACSA.2024.0151043.

[12] Ezziyyani, M., Cherrat, L., Jebari, H., Rekiek, S., & ahmed, N.a. "CNN-Based Plant Disease Detection: A Pathway to Sustainable Agriculture." In: Ezziyyani, M., Kacprzyk, J., Balas, V.E. (eds.) Proc. Int. Conf. Adv. Intell. Syst. Sustain. Dev. (AI2SD'2024). LNNS, vol. 1403, pp. 679–696. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-91337-2_62.

[13] Ezziyyani, M., Cherrat, L., Rekiek, S., & Jebari, H. "Image Classification of Moroccan Cultural Trademarks." In: Ezziyyani, M., Kacprzyk, J., Balas, V.E. (eds.) Proc. Int. Conf. Adv. Intell. Syst. Sustain. Dev. (AI2SD'2024). LNNS, vol. 1403, pp. 767–779. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-91337-2_68.

[14] Gouiza, N., Jebari, H., & Reklaoui, K. "IoT in smart farming: A review." In: Ezziyyani, M., Kacprzyk, J., Balas, V.E. (eds.) AI2SD 2023, Lecture Notes in Networks and Systems, vol. 930, pp. 142–153. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-54327-2_11.

[15] Gouiza, N., Jebari, H., & Reklaoui, K. "IoT in Agriculture: Use Cases and Challenges." In: Ezziyyani, M., Kacprzyk, J., Balas, V.E. (eds.) Proc. Int. Conf. Adv. Intell. Syst. Sustain. Dev. (AI2SD'2024). LNNS, vol. 1402, pp. 491–505. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-91334-1_42.

[16] Jardine, A. K., Lin, D., & Banjevic, D. "A review on machinery diagnostics and prognostics implementing condition-based maintenance." Mechanical systems and signal processing, 20(7), 1483-1510, 2006. https://doi.org/10.1016/j.ymssp.2005.09.012.

[17] Heng, A., Zhang, S., Tan, A. C., & Mathew, J. "Rotating machinery prognostics: State of the art, challenges and opportunities." Mechanical systems and signal processing, 23(3), 724-739, 2009. https://doi.org/10.1016/j.ymssp.2008.06.009.

[18] Si, X. S., Wang, W., Hu, C. H., & Zhou, D. H. "Remaining useful life estimation–a review on the statistical data driven approaches." European journal of operational research, 213(1), 1-14, 2011. https://doi.org/10.1016/j.ejor.2010.11.018.

[19] Widodo, A., & Yang, B. S. "Support vector machine in machine condition monitoring and fault diagnosis." Mechanical systems and signal processing, 21(6), 2560-2574, 2007. https://doi.org/10.1016/j.ymssp.2006.12.007.

[20] Breiman, L. "Random forests." Machine learning, 45(1), 5-32, 2001. https://doi.org/10.1023/A:1010933404324.

[21] Rasmussen, C. E., & Williams, C. K. I. "Gaussian processes for machine learning." MIT Press, Cambridge, MA, 2006.

[22] Lei, Y., Jia, F., Lin, J., Xing, S., & Ding, S. X. "An intelligent fault diagnosis method using unsupervised feature learning towards mechanical big data." IEEE Transactions on Industrial Electronics, 63(5), 3137-3147, 2016. https://doi.org/10.1109/TIE.2016.2519325.

[23] Janssens, O., Slavkovikj, V., Vervisch, B., Stockman, K., Loccufier, M., Verstockt, S., ... & Van de Walle, R. "Convolutional neural network based fault detection for rotating machinery." Journal of Sound and Vibration, 377, 331-345, 2016. https://doi.org/10.1016/j.jsv.2016.05.027.

[24] Malhotra, P., TV, V., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P., & Shroff, G. "Multi-sensor prognostics using an unsupervised health index based on LSTM encoder-decoder." arXiv preprint arXiv:1608.06154, 2016.

[25] Zhu, J., Chen, N., & Peng, W. "Estimation of bearing remaining useful life based on multiscale convolutional neural network." IEEE Transactions on Industrial Electronics, 66(4), 3208-3216, 2018. https://doi.org/10.1109/TIE.2018.2844856.

[26] Wu, D., Jennings, C., Terpenny, J., Gao, R. X., & Kumara, S. "A comparative study on machine learning algorithms for smart manufacturing: tool wear prediction using random forests." Journal of Manufacturing Science and Engineering, 139(7), 2017. https://doi.org/10.1115/1.4036350.

[27] Khumprom, P., & Yodo, N. "A data-driven predictive prognostic model for lithium-ion batteries based on a deep learning algorithm." Energies, 12(4), 660, 2019. https://doi.org/10.3390/en12040660.

[28] Jebari, H., Mechkouri, M.H., Rekiek, S., & Reklaoui, K. "Poultry-edge-AI-IoT system for real-time monitoring and predicting by using artificial intelligence." International Journal of Interactive Mobile Technologies, 17(12), 58–70 (2023). https://doi.org/10.3991/ijim.v17i12.38095.

[29] Gouiza, N., Jebari, H., & Reklaoui, K. "Integration for IoT-enabled technologies and artificial intelligence in diverse domains: Recent advancements and future trends." Journal of Theoretical and Applied Information Technology, 102(5), 1975–2029 (2024).

[30] Nectoux, P., Gouriveau, R., Medjaher, K., Ramasso, E., Chebel-Morello, B., Zerhouni, N., & Varnier, C. "PRONOSTIA: An experimental platform for bearings accelerated degradation tests." IEEE International Conference on Prognostics and Health Management, 1-8, 2012. https://doi.org/10.1109/ICPHM.2012.6299546.

[31] Kingma, D. P., & Ba, J. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980, 2014.

[32] Saxena, A., Goebel, K., Simon, D., & Eklund, N. "Damage propagation modeling for aircraft engine run-to-failure simulation." 2008 International Conference on Prognostics and Health Management, 1-9, 2008. https://doi.org/10.1109/PHM.2008.4711414.

[33] Jebari, H., Rekiek, S., Ezziyyani, M., & Cherrat, L. "Artificial intelligence for optimizing livestock management and enhancing animal welfare." In: Ezziyyani, M., Kacprzyk, J., Balas, V.E. (eds.) Proc. Int. Conf. Adv. Intell. Syst. Sustain. Dev. (AI2SD'2024). LNNS, vol. 1403, pp. 790–800. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-91337-2_70.

[34] Zhang, A., Li, S., Cui, Y., Yang, W., Dong, R., & Hu, J. "Limited data rolling bearing fault diagnosis with few-shot learning." IEEE Access, 7, 110895-110904, 2019. https://doi.org/10.1109/ACCESS.2019.2934233.

[35] Wilkinson, M., Darnell, B., Van Delft, T., & Harman, K. "Comparison of methods for wind turbine condition monitoring with SCADA data." IET Renewable Power Generation, 8(4), 390-397, 2014. https://doi.org/10.1049/iet-rpg.2013.0187.

[36] Astolfi, D., Castellani, F., & Terzi, L. "Fault prevention and diagnosis through SCADA temperature data analysis of an onshore wind farm." Diagnostyka, 15(2), 2014.

Federated Machine Learning for Monitoring Student Mental Health in Kazakhstan

Bakirova Gulnaz[1], Bektemyssova Gulnara[2], Nor'ashikin Binti Ali[3]
Department of Computer Engineering, International University of Information Technology, Kazakhstan [1, 2]
College of Computing and Informatics, Universiti Tenaga Nasional, Kajang, Selangor, Malaysia [3]

***Abstract*—Federated Learning (FL) offers a privacy-preserving and decentralized paradigm for machine learning, making it particularly suitable for analyzing sensitive psychological and physiological data. This study aims to develop and evaluate a federated learning framework for assessing the psycho-emotional well-being of students in Kazakhstani educational institutions, where data privacy and infrastructural constraints pose significant challenges. We benchmark three FL algorithms, such as FedAvg, FedOpt, and FedProx, on heterogeneous, institution-level datasets that combine sleep, dietary, activity, and self-reported emotional measures. Experiments simulate cross-device, non-IID deployments and evaluate convergence, accuracy, and stability across ten communication rounds. Results show that FedProx attains the best trade-off between accuracy and stability under non-IID conditions (peak accuracy is 99.9%), while FedOpt provides faster early convergence, and FedAvg performs well for more homogeneous partitions. The methodological contribution comprises optimized aggregation and adaptive client weighting to mitigate non-IID effects in resource-constrained educational settings. These findings validate FL as a scalable, privacy-preserving approach for mental health monitoring in education and support its use for early intervention and resilience tracking. The proposed framework contributes to data-driven mental health policy design in educational systems, addressing both ethical and infrastructural considerations. The study discusses limitations of the simulated setup and outlines directions for broader deployment and cross-silo validation.**

Keywords*—Federated Learning; data privacy; FedOpt; FedAvg; FedProx; mental health; non-IID data; educational data mining; psychological analytics

I. Introduction

Machine learning (ML), a subset of artificial intelligence (AI), has experienced rapid advancements, enabling powerful computational techniques to analyze and interpret vast amounts of data. Unlike static, rule-based systems, ML algorithms learn from data, identifying patterns and improving performance over time. This adaptability has facilitated breakthroughs in areas such as image recognition, natural language processing, recommendation systems, and predictive analytics [1]. Central to this progress is the availability of big data, characterized by its volume, velocity, and variety. These datasets, encompassing sources like IoT sensor readings and social media interactions, form the foundation for training complex models capable of delivering highly accurate insights [2].

Despite its potential, traditional ML methods often rely on centralized data collection, raising significant privacy and security concerns. Sensitive data aggregated on centralized servers is vulnerable to breaches and may conflict with regulations like the GDPR [3] and CCPA. These limitations have catalyzed the emergence of FL, a decentralized machine learning paradigm. FL enables collaborative model training across distributed devices such as smartphones, IoT nodes, and edge computing systems while keeping raw data localized. Instead of transferring data to a central server, each device trains a local model and transmits only model updates (e.g., gradients or weights) to a central server for aggregation. By preserving data privacy and minimizing communication overhead, FL has gained prominence in sensitive domains such as healthcare, education, and finance [4].

In this study, we focus on the analysis for monitoring student mental health in Kazakhstan. Applying the FL mental health [5] analysis presents unique challenges due to the country's diverse cultural, social, and educational contexts. Beyond behavioral and physiological differences, students' emotional well-being is influenced by various factors, such as academic workload, which includes lessons, homework, and the preparation for and completion of exams. These stressors contribute to emotional burnout and introduce additional variability into the data, complicating the analysis and modeling process [6].

Given these challenges, this study seeks to develop and evaluate a federated learning framework for analyzing the psycho-emotional well-being of students in Kazakhstan. By integrating privacy-preserving algorithms such as FedAvg, FedOpt, and FedProx, the proposed framework addresses critical challenges associated with data heterogeneity, infrastructural limitations, and confidentiality within educational environments.

This study is organized as follows: Section II provides a review of related studies on federated learning algorithms and their applications in privacy-preserving educational analytics. Section III outlines the dataset, data preprocessing methods, and experimental design, detailing the implementation of FedAvg, FedOpt, and FedProx. Section IV presents experimental results and offers a comparative evaluation of algorithm performance under heterogeneous data conditions. It also discusses the implications of the findings for enhancing psycho-emotional well-being monitoring in educational contexts. Section V concludes the study by summarizing key findings and outlining future research directions.

In summary, this study investigates how federated learning can enhance the monitoring of students' mental health in

Kazakhstan while preserving data privacy and addressing infrastructural limitations. This framework aims to demonstrate the feasibility of applying FL under heterogeneous educational conditions and to contribute to the development of ethical, data-driven mental health monitoring systems in the Kazakhstani context.

II. Related Work

The specific key challenges are as follows as data privacy and security. The analysis of psycho-emotional data involves sensitive personal information, including sleep patterns, dietary habits, physical activity levels, and emotional responses. Ensuring the privacy and security of this data is paramount to protect students' confidentiality. Data Heterogeneity Students' behavioral data is highly non-IID (non-independent and identically distributed) [7]. Emotional Burnout and Stress Academic-related stressors, such as homework, lesson loads, and exam preparation, exacerbate emotional burnout among students. These factors not only impact students' psycho-emotional states but also introduce temporal variability in the data. As a result, FL models must be adaptable to dynamic changes in students' emotional states over time. In our work, we have considered algorithms such as FedAvg, a better-known efficient algorithm for training machine learning models on multiple clients in a privacy-preserving manner. Traditionally, FedAvg uses a central server to distribute parameters to clients and aggregate updates, but it has difficulties with privacy and communication efficiency. FedOpt achieves faster convergence, lower communication overhead, and higher accuracy with fewer training epochs compared to benchmarks. Research objectives are: developing and evaluating privacy-preserving FL algorithms (e.g., FedAVG, FedOPT, FedProx) tailored to psycho-emotional state analysis. Address data heterogeneity through techniques such as regularization, clustering-based FL, or adaptive weighting of client updates. Optimize FL frameworks for resource-constrained devices by implementing efficient communication strategies (e.g., gradient sparsification, local updates). Validate the proposed framework using real-world datasets that include students' behavioral and psycho-emotional metrics.

A. Revised Research Contributions

1) Development of a localized Federated Learning (FL) framework specifically designed for psycho-emotional data analysis in educational settings. The proposed framework integrates cross-device horizontal FL with synchronous aggregation, optimized for limited network connectivity and heterogeneous client environments typical of Kazakhstani institutions.

2) Algorithmic enhancement of FedAvg, FedOpt, and FedProx through adaptive weighting and proximal regularization to mitigate non-IID effects. The study introduces modified update strategies that dynamically adjust client contributions based on data volume and distribution, improving global model convergence and stability compared to baseline implementations.

3) Integration of behavioral and emotional indicators into federated model evaluation, linking technical model performance with psycho-emotional metrics such as burnout and well-being. This enables data-driven validation of FL's applicability for educational mental health monitoring, and supports model interpretability through domain-relevant insights.

In FL, the global model is trained iteratively over several rounds of client-server interaction, without centralized data collection, where each interaction is defined as an FL round. This process continues until the model reaches the desired level of accuracy [8]. Fig. 1 shows the architecture of federated learning that was applied in our work. It takes several rounds to update the model to the global model [9]. FL approaches can be categorized into three types: synchronous, asynchronous, and semi-asynchronous. In our work, we propose synchronous FL, in which the parameter server must wait for clients, which leads to noticeable waiting time due to edge heterogeneity. It should be considered that the heterogeneity of edge nodes means that they have different computational and communication capabilities. Although this method introduces waiting time due to edge heterogeneity, it provides better model convergence and accuracy, making it suitable for sensitive applications like psycho-emotional state analysis in education.

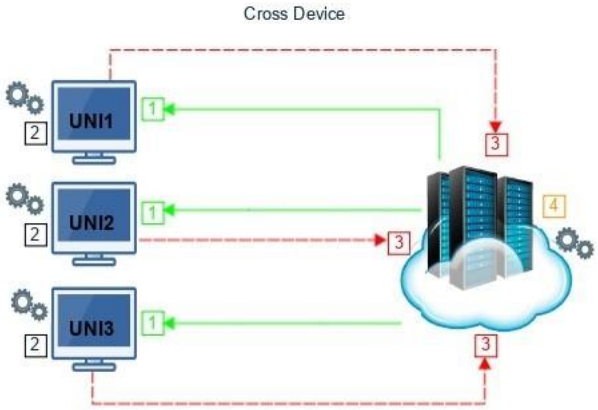


Fig. 1. The architecture of federated learning.

FL approaches can be categorized into three types by learning methods, architectures, and aggregation strategies, including synchronous FL, where the server waits for all clients to complete training before aggregating updates. While this method ensures consistency and high accuracy, it may lead to inefficiencies due to edge heterogeneity—variations in computational and communication capabilities among clients [10]. In asynchronous FL, clients communicate updates independently, reducing waiting times but potentially introducing inconsistencies in global model updates [11]. Semi-Asynchronous FL is a hybrid approach that balances the benefits of synchronous and asynchronous methods.

Experiments using synchronous method in FL [12], where clients must download global model updates at one point in time, and the server waits for clients to complete the training tasks. This synchronous optimization mechanism causes clients with limited network or insufficient hardware resources to lag; the server and other clients have to wait, and the learning efficiency decreases, i.e., there is an overload effect, but for our problem, we paid attention to the obvious advantages of this

synchronous update method in that the model converges easily and has high accuracy.

FL [13, 14] is a machine learning technique that focuses on mutual learning of a model by multiple clients, where the data remains decentralized. FL is a critical technology in the current era of big data and artificial intelligence, recognized for its critical role in protecting data privacy and eliminating the need to transfer and process huge amounts of data. FL allows the benefits of machines to remain intact. Unlike the traditional centralized learning approach, it facilitates collaborative training of statistical models sharing parameter updates instead of raw data [15, 16]. In our work, we have considered algorithms such as FedAvg, a better-known efficient algorithm for training machine learning models on multiple clients in a privacy-preserving manner. Traditionally, FedAvg uses a central server to distribute parameters to clients and aggregate updates, but has difficulties with privacy and communication efficiency [17]. FedOpt achieves faster convergence, lower communication overhead [18], and higher accuracy with fewer training epochs compared to benchmarks [19].

In [20], the authors present a comprehensive scheme for FL provisioning under time constraints in IoT environments using push-pull communication mode. By combining utility-based scheduling and efficient client selection, higher accuracy and lower latency are achieved compared to traditional methods. The approach is particularly effective in heterogeneous and resource-constrained networks, paving the way for practical implementation of FL in IoT and edge computing. In [21], ClipFL, a novel method for handling noisy labels in Federated Learning by identifying and excluding noisy clients, is discussed. The method demonstrates high accuracy, fast convergence, and reduced communication overhead, making it a practical solution for real FL systems with noisy data. The simplicity and scalability of the method make it suitable for large-scale deployment even in resource-constrained and heterogeneous environments. Traditional federated learning (FL) methods are enhanced, introducing a split learning approach where client-side and server-side models are trained together, which significantly improves the performance and efficiency of the interaction [22].

In [23], the importance of robustness in federated learning is emphasized, privacy and security concerns are addressed, and practical security mechanisms are described. By addressing these vulnerabilities, FL can be effectively deployed across industries, ensuring data integrity and model robustness in distributed environments. In the future, dynamic, scalable, and lightweight solutions should be prioritized to make FL both secure and efficient. One of the studies [24] shows that L2GDV (Loopless Local Gradient Descent with Varying step size) significantly improves the efficiency and performance of fuzzy learning by combining adaptive regularization step size methods. It provides robust convergence guarantees while reducing communication costs, which makes it suitable for real FL applications with heterogeneous data and limited resources. The study [25] emphasizes the vulnerability of FL to data poisoning attacks using unwanted samples, especially when using robust aggregation schemes such as Krum and Bulyan. While aggregation algorithms mitigate some attacks, they remain insufficient against adaptive adversary techniques.

Recent works [26–27] have explored centralized machine learning approaches for mental health prediction and early intervention, demonstrating promising results in healthcare and behavioral analytics. However, few studies have applied federated frameworks in educational contexts, particularly within developing regions, where data heterogeneity and privacy concerns remain major obstacles. However, these studies rely on centralized architectures, whereas our work introduces a federated paradigm suitable for privacy-constrained educational environments.

III. Proposed Methodology

FL utilizes synchronous, asynchronous, and semi-synchronous training methods, each suited to different network conditions. Synchronous learning ensures high accuracy by aggregating updates only after all clients submit them, but it requires stable connectivity and is sensitive to slow clients. Asynchronous learning allows clients to update independently, accelerating training in heterogeneous environments but risking model instability. Semi-synchronous learning balances these approaches by waiting for a subset of clients (e.g., 80%) before updating, reducing delays while maintaining stability. Given these factors, synchronous learning was chosen to ensure consistent model updates.

Although FL is classified into horizontal and vertical types, where horizontal FL involves similar features across different users, while vertical FL integrates different features for the same users. FL can also be Cross-Silo, applied in large organizations with stable connections, or Cross-Device, involving numerous heterogeneous personal devices with limited resources. Our study employs synchronous learning, horizontal FL, and Cross-Device FL for optimized distributed training. The structure of FL is illustrated in Fig. 2.

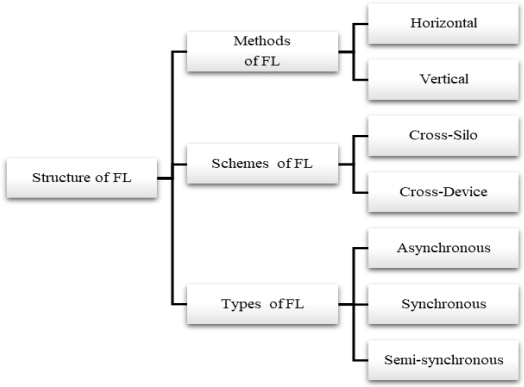


Fig. 2. Structure of FL.

A. Description of Method

These methods are used to predict students' psychological and emotional state based on nutrition, activity, and sleep data, such as interaction type. The distribution of the data presented below: Type interactions: When solving our task, we choose the Cross-Device type of interaction, since the task involves collecting data from devices where data is stored locally. The data are heterogeneous in terms of quality and distribution (different modes of sleep, nutrition, activity of students). Devices may be resource constrained (low processing power,

variable network). Cross-Device FL performs well in tasks where devices are limited in computational resources, and the data have a high level of heterogeneity. Selection: Horizontal. Distribution data: Each device has a similar data structure: sleep metrics, nutrition, activity, but the data belongs to different users. Horizontal training allows effective work with such scenarios, as it focuses on combining knowledge from similar data structures. Selection: synchronous. In our scenario, we will be able to organize synchronous student connection devices. Synchronous learning allows the server to produce an immediate update of the global model, which gives a good result for accuracy.

B. Selection of FL Algorithms

In addition, to effectively handle data heterogeneity in non-IID student datasets, we selected federated optimization (FO) algorithms that adapt to variability: FedAvg serves as a baseline approach, which can be enhanced by assigning different weights to clients with varying data distributions. FedProx incorporates regularization to mitigate the effects of dominant local updates, improving convergence in heterogeneous environments. FedOpt employs adaptive step sizes for both client and global models, facilitating faster convergence and enhanced scalability for complex, large-scale datasets.

FedAvg (Federated Averaging) [28] is a standard algorithm for federated learning that provides distributive learning on local devices, minimizing data exchange with the server. The algorithm works as follows: each client trains its model on local data and sends updates to the server. The server aggregates all local updates, computing an average for the model parameters. This method is repeated till convergence is achieved. The basic Formula (1):

$$\omega^{t+1} = \frac{1}{K}\sum_{k=1}^{K} \omega_k^{t+1} \quad (1)$$

where, ω_k^{t+1} updated model parameters from customer k, and K - number of customers.

FedOpt (Federated Optimization) [29] is an improvement to the FedAvg algorithm that incorporates the use of adaptive optimizers, such as Adam, for local model updates. Instead of using simple gradient descent as in FedAvg, FedOpt applies optimizers for each client, which helps to speed up convergence and reduce the impact of local errors. FedOpt can accommodate different learning rates for each client, making it more flexible and efficient. The basic Formula (2):

$$\omega^{t+1} = \omega_k^t - \mu_k \nabla L_k(\omega_k^t) \quad (2)$$

where, μ_k is the learning rate for client k, and ∇L_k is the gradient of the loss function for the client.

FedProx (Federated Proximal Optimization) [30] is an enhancement to the FedAvg algorithm designed to deal with heterogeneous data. It adds proximal regularization to minimize the impact of local differences in the data and improve learning stability. Instead of simply weighing local updates as in FedAvg, FedProx accounts for deviations of the local model from the global model using a proximal term that helps improve convergence and avoid overfitting. The basic Formula (3):

$$\omega_k^{t+1} = \omega_k^t - \mu_k \nabla L_k(\omega_k^t) - \mu(\omega_k^t - \omega) \quad (3)$$

where, μ is the regularization coefficient, and ω are the global parameters of the model.

The advantages of the algorithms are FedAVG - Easy implementation and low requirements for computing resources. FedOPT - use of adaptive optimizers accelerates convergence and helps avoid overfitting on unstable data. FedPROX is suitable for cases with heterogeneous data and high variability between clients. Proximal regularization helps to improve convergence.

The limitations of the algorithms are FedAVG - sensitivity to data heterogeneity, where data on different devices vary significantly, which can slow down the learning and degrade the quality of the model. FedOPT - Increased computational costs on the client, as more sophisticated optimization techniques must be used. FedPROX - requires tuning of the regularization parameter, which may add complexity to the optimization process.

Algorithms like FedSGD, FedAvg, FedProx, SCAFFOLD, and FedOpt were considered for this study, but FedAvg, FedProx, and FedOpt were selected for further study. It is shown in Table I.

TABLE I DETAILS OF ALGORITHMS

Method	Non-IID	Convergence rate	Flexibility to heterogeneity of clients	Use of server optimizers
FedAvg	Bad	Slowly	No	No
FedProx	Medium	Medium	Yes	No
FedOpt	Good	Fast	Yes	Yes

In summary, FedAvg is a basic algorithm for federated learning, effectively applicable if the data on the clients is similar. FedOpt improves FedAvg using adaptive optimizers, which accelerates convergence and improves model accuracy. FedProx focuses on heterogeneous data and incorporates proximal regularization, which improves training stability in the face of large differences between clients. Each of these methods has its own features and applications, and the choice of the appropriate algorithm depends on the data structure and computational resources.

In federated model training, FedAvg, FedOpt, and FedProx optimize learning on distributed clients while minimizing communication with the central server. FedAvg is effective when client data distributions are similar, offering a computationally efficient solution for homogeneous datasets, such as student learning or physical activity data. FedOpt adapts to heterogeneous data using adaptive optimizers (e.g., Adam), improving convergence speed and model accuracy, making it suitable for dynamic environments. FedProx addresses highly heterogeneous data by incorporating regularization, mitigating localized errors, and enhancing model stability and accuracy in diverse student datasets. In conclusion, FedAvg will be a simple and efficient solution when the data is homogeneous between clients, in case of heterogeneous data and different device types, we recommend using FedOpt or FedProx to improve model quality and convergence. Provided the problem requires regularization and consideration of local differences, FedProx will provide a stable solution. Thus, the choice of algorithm depends on the nature of the data and the goals: speed of

training and accuracy of the model or the ability to adapt to differences in customer data.

C. Experimental Settings

The tasks are focused on client-side data processing, where local models are applied directly on the client devices. Sending updates to the server. Server-side update processing, creating a global model. Sending global models to customers. Global model with client data is processed on the client side. These steps will ensure stability, privacy, and efficiency in implementing the study in real-world use cases. For our heterogeneous data problem, the choice of methods and approaches must consider the heterogeneous distribution of data across devices.

This study employs a client-side learning approach, ensuring that raw data remains on user devices, thus preserving privacy while still enabling effective global model updates. The following stages describe each step of this process in detail.

Stage 1: Data Preparation

At this stage, data is collected through a web platform where participants complete the Maslach Burnout Inventory (MBI) questionnaire. Once collected, the data is cleaned by removing duplicates, input errors, and outdated records, then normalized and standardized. Special attention is given to simulating real-world conditions with heterogeneous (non-IID) data to enhance model robustness and realism.

Stage 2: Local Training on Clients

Each client trains a random forest model locally on their own data without transmitting raw information to the server. The updated model parameters (e.g., tree weights and performance metrics) are sent to the central server to ensure data privacy. This approach allows the model to adapt to the unique characteristics of each client's dataset and improves prediction accuracy.

Stage 3: Model Aggregation on the Server

The server collects updates from all clients and aggregates them using algorithms such as FedAvg, FedProx, or FedOpt. Aggregation is weighted according to the volume of data on each client to ensure fair contribution. After aggregation, the server evaluates model convergence using metrics like MSE and R^2 and adjusts training parameters if needed.

Stage 4: Model Updating and Client Communication

The updated global model is sent back to clients via secure communication channels. Clients integrate the new model into their local environment and resume training on their data. This iterative process of model distribution and retraining ensures continuous improvement and adaptation to new data.

Stage 5: Model Evaluation

After several rounds of training, the model is evaluated using both classification (Accuracy, Precision, Recall, F1-score) and regression (MSE, R^2) metrics. The impact of data heterogeneity among clients is analyzed to assess its effect on model accuracy and convergence. This comprehensive evaluation helps identify weaknesses and optimize aggregation strategies and learning parameters. Fig. 3 illustrates these steps.

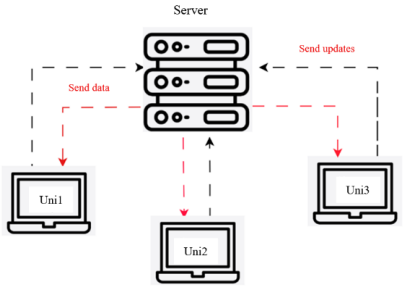
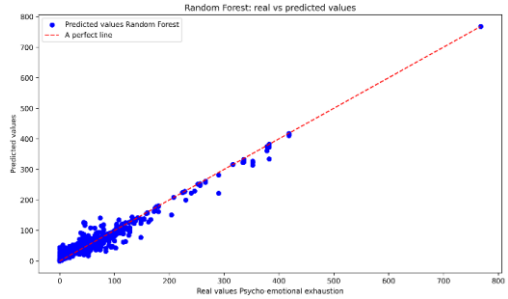


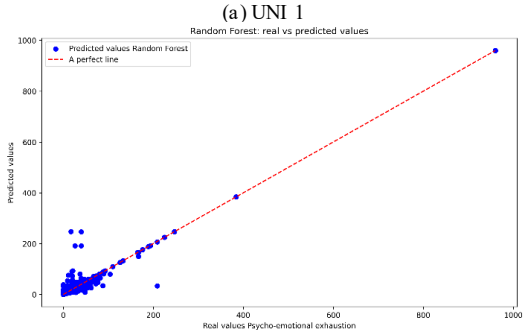
Fig. 3. Federated learning architecture: client-server interaction.

The Random Forest method was applied to the local model; this algorithm was determined by conducting experiments during which methods such as logical regression, linear regression, and time series were considered. The Random Forest method is an ensemble machine learning algorithm that uses multiple decision trees to solve classification and regression problems. The essence of the method is to train multiple trees on a Random Forest of training data and features, which improves the generalization ability of the model and reduces overfitting. Each tree is trained on a random sample of data with return (bootstrap method), and for classification, the final decision is made by a majority vote of the trees, and for regression, through averaging of predictions. Before sending them to the server after training on local devices, the following results were obtained, which are presented in Fig. 4.

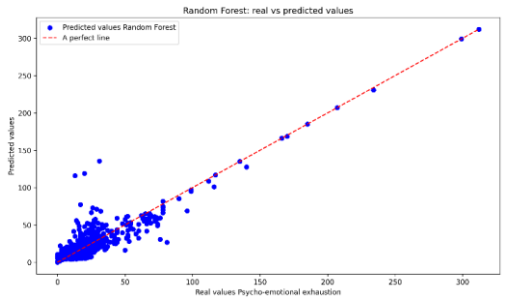
The graphs' description illustrates the relationship between actual and predicted values of Psycho-emotional exhaustion. The horizontal axis represents actual values, while the vertical axis shows predicted values. A red dashed line indicates a perfect match. For low and medium values, predictions closely align with the ideal line, demonstrating high accuracy. As values increase, a slight deviation appears, with greater scatter at higher levels (~1000). Despite this, predictions remain well-correlated with the actual data, maintaining stable performance across different ranges.



(a) UNI 1



(b) UNI 2



(c) UNI 3

Fig. 4. Comparison of the results of the random forest method in (a) UNI 1, (b) UNI 2, (c) UNI 3.

The pros of the method are high accuracy, especially for complex and high-dimensional data, robustness to overfitting due to the ensemble approach, handling of missing values, and estimation of feature importance, which aids in model interpretation. The cons are low interpretability compared to single trees, difficulties with prediction on very large data due to the large number of trees. Random forest is widely used in classification and regression tasks such as disease diagnosis, credit scoring, price prediction, and tasks requiring feature selection. In the federated learning task, random forest is an effective method for combining models from different clients, providing robustness to local data and improving overall performance. After numerous experiments on the server side, a local model was picked up and sent to the client side, where training began. The graphs compare the actual values and predicted values of the Random Forest model for the Psycho-emotional exhaustion index. For this purpose, data from different devices were used, which allows us to clearly demonstrate the effectiveness of the model in the conditions of heterogeneous data. The Random Forest model shows a good ability to predict Psycho-emotional exhaustion, especially in the small and medium ranges of values. Visualization confirms that the model successfully captures the data patterns, although some deviation from the ideal line is observed at higher values (see Table II).

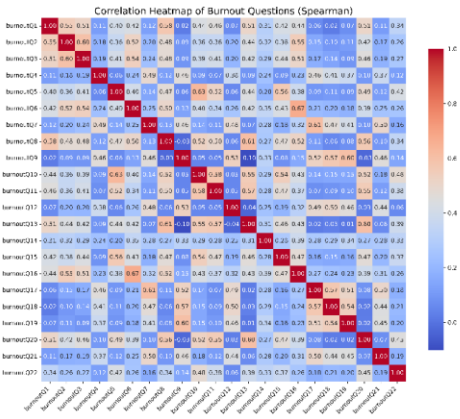
TABLE II ACCURACY BEFORE SENDING IT TO THE SERVER

	UNI 1	UNI 2	UNI 3
Accuracy	0.915676	0.554274	0.858384

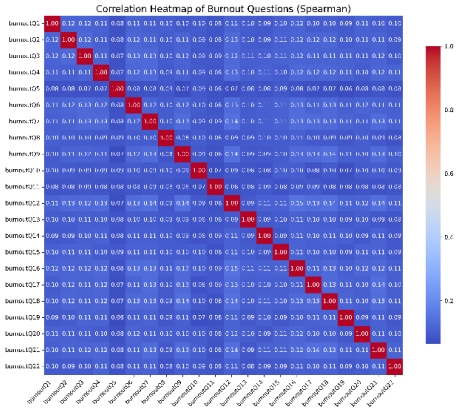
1) Description of the preparatory process: To solve the task, a website was created where the data collection of each individual student on their personal devices was organized. For example, it was decided to make a psycho-emotional and physical analysis of the state of students and schoolchildren based on the collected database. Currently, the FL has not yet started its march in Kazakhstan, so my work is of an introductory nature on the possibilities of using the FL. The purpose of the experiment is to make a practical demonstration of the whole process of organizing FL steps on a specific example described in the study.

Three educational institutions participated in the experiment; data were collected from the beginning of the learning process, from September 1, for 2 months before the first milestone control. A website was created for data collection, where all data were stored in a database based on the educational institution. The aim was to analyze the mental and emotional state of students from the moment of the beginning of the study and by the end of the first milestone control to determine how much their state had changed by the time of passing the first milestone control.

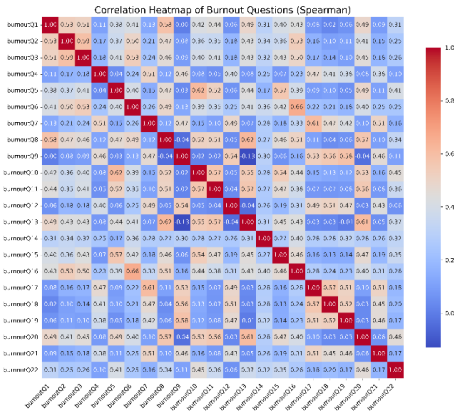
These heatmaps provide an overview of how different questions about burnout relate to each other, which can be useful for identifying patterns and determining which questions may share common underlying factors or experiences, as in Fig. 5.



(a) UNI 1



(b) UNI 2



(c) UNI 3

Fig. 5. Correlation heatmap of burnout questions in (a) UNI 1, (b) UNI 2, (c) UNI 3.

The heatmaps display Spearman correlation coefficients for burnout-related questions (burnoutQ1–burnoutQ22), with values ranging from -1 to 1. Red indicates strong positive correlations, blue represents weak or no correlations, and white signifies no correlation (0). Heatmap 1 highlights strong correlations between certain burnout questions. Heatmap 2 shows generally lower correlation values, with many closes to zero. Heatmap 3 reveals some strong correlations (e.g., burnoutQ1–Q3, burnoutQ6–Q9), but overall, the values remain weaker. These heat maps help identify patterns and relationships between burnout factors.

Application of FL was aimed at predicting further psycho-emotional and physical state of students for the next boundary control and readiness to pass the winter session. The heterogeneity of the data can be determined based on the following characteristics shown in Table III.

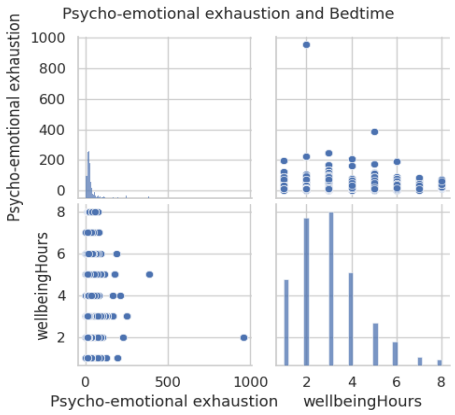
TABLE III CHARACTERISTICS OF THE DATA

Data type	Heterogeneous	Homogeneous
Nutrition	'Breakfast', 'Dinner', 'Lunch', 'total_meals'.	
Physical activity	'intensity', 'activityType', 'duration'	
Sleep and rest	'wellbeingHours0', 'wellbeingHours0', 'wellbeingHours1'.	
Emotional state	'Psycho-emotional exhaustion', 'burnoutQ1'- 'Q22'.	
Identifiers and meta-data		'student_id', 'entry_date'

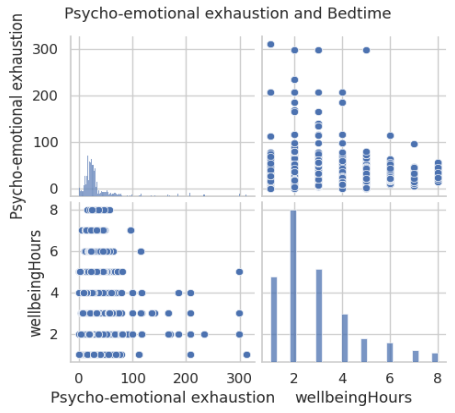
2) Heterogeneity analysis: Students' data exhibits heterogeneity due to variations in behavioral patterns, including nutrition, physical activity, sleep, and emotional state. For instance, dietary habits differ among students, with some skipping meals while others follow a structured diet. Similarly, physical activity levels vary, ranging from active sports participation to a sedentary lifestyle. Emotional burnout scores also fluctuate based on individual stress levels and coping mechanisms. Conversely, homogeneous attributes include 'entry date' (consistent format across devices) and 'student_id' (a unique identifier that does not impact target metrics). These attributes remain stable regardless of individual behaviors in Fig. 6. To address data heterogeneity, preprocessing is required to normalize activity and nutrition data, ensuring consistency across samples. Additionally, leveraging homogeneous attributes (e.g., timestamps, unique identifiers) aids in standardization and enhances model reliability.

The pair plots illustrate a weak negative correlation between Psycho-emotional exhaustion and wellbeingHours. Most students show low sleep duration regardless of exhaustion level, indicating that increased rest does not consistently reduce emotional fatigue. The presence of outliers reflects the heterogeneity of behavioral data among students.

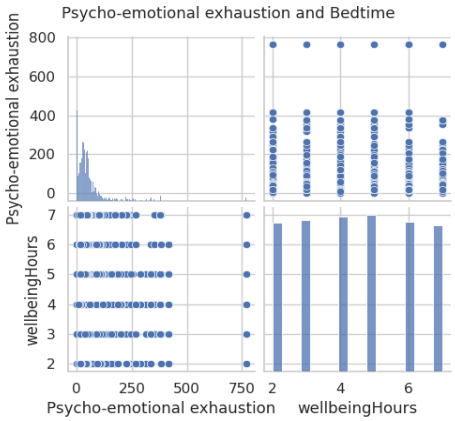
The following results highlight the comparative performance and convergence behavior of the selected FL algorithms under non-IID educational data conditions.



(a) UNI 1



(b) UNI 2



(c) UNI 3

Fig. 6. Fluctuations in psycho-emotional exhaustion over one month in (a) UNI 1, (b) UNI 2, (c) UNI 3.

IV. RESULTS AND DISCUSSION

In this study, the performance of three federated learning algorithms, namely FedAVG, FedOPT, and FedProx, was benchmarked on a classification task aimed at predicting students' psycho-emotional states. The primary evaluation metric was Accuracy, measured on test data after each of ten training rounds. Table IV and Fig. 7 illustrate the dynamics of model accuracy across all rounds, providing insight into the learning stability and convergence behavior of each algorithm.

TABLE IV ACCURACY DYNAMICS FOR ALGORITHMS

Rounds	Accuracy		
	FedAVG	FedOpt	FedProx
1	0.793071234	0.8383071568	0.9682882882
2	0.850712345	0.8707123457	0.9793333333
3	0.9366666666	0.942457545	0.989333333
4	0.95956756756	0.9895675675	0.9866633333
5	0.97666666666	0.926666666	0.999936
6	0.9956756756	0.9695675675	0.99657878888
7	0.97666666666	0.9876666666	0.988888999
8	0.96666666666	0.9866666666	0.9968288288
9	0.995675675	0.995675675	0.9933333333
10	0.999675699	0.999675699	0.998888999

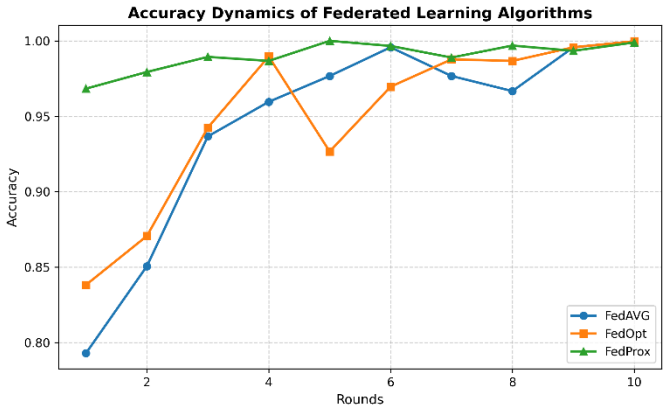


Fig. 7. Accuracy dynamics for algorithms.

As shown in Fig. 7, all three algorithms exhibit an upward trend in accuracy as training progresses, demonstrating the effectiveness of the federated learning approach in distributed educational data environments. Among them, FedProx consistently achieved the highest accuracy, particularly in later rounds where it reached 99.99%, indicating its robustness in managing heterogeneous, non-IID data. This superior performance is attributed to its proximal regularization, which stabilizes updates and reduces divergence caused by local data variability across clients:

a) FedOpt, which incorporates adaptive optimizers (e.g., Adam), also demonstrated strong convergence characteristics, achieving accuracy levels close to FedProx while maintaining lower communication overhead. However, its slightly lower accuracy in the final rounds suggests that additional regularization could further enhance its performance under highly heterogeneous conditions.

b) FedAVG, while computationally efficient, displayed slower convergence and lower final accuracy compared to the other two methods. Its performance plateaued in the later rounds, emphasizing its sensitivity to client data heterogeneity. Nonetheless, FedAVG remains a practical solution for homogeneous datasets or resource-constrained environments, where computational simplicity is prioritized over precision.

c) Overall, the results confirm that FedProx is the most effective algorithm for addressing data heterogeneity and achieving high model accuracy in psycho-emotional state prediction. FedOpt offers a balanced trade-off between speed and accuracy, whereas FedAVG is better suited for uniform data distributions.

These findings support the study's objectives by demonstrating that federated learning algorithms can be effectively applied to privacy-preserving educational analytics. The results offer actionable insights for educational institutions, researchers, and policymakers, emphasizing the potential of federated learning to enhance the monitoring of students' psycho-emotional well-being while ensuring data confidentiality and accommodating diverse learning environments across regions.

Moreover, the results directly align with the study's objective of developing a federated learning framework for monitoring mental health in Kazakhstan, confirming that such systems can be implemented effectively under heterogeneous educational conditions.

V. CONCLUSION AND FUTURE SCOPE

This study aimed to develop and evaluate a privacy-preserving federated learning framework for analyzing the psycho-emotional well-being of students in Kazakhstan, particularly under conditions of data heterogeneity and non-IID distributions. To achieve this, three federated learning algorithms, FedAvg, FedOpt, and FedProx, were benchmarked across ten communication rounds in a classification task. The results revealed distinct performance dynamics, convergence rates, and stability patterns under federated conditions.

FedAvg showed steady improvement in accuracy from 0.793 in round 1 to 0.9997 in round 10, indicating reliable convergence despite minor fluctuations in intermediate stages. FedOpt achieved slightly higher accuracy in the early rounds (0.9425 and 0.9895 in rounds 3 and 4) and demonstrated faster convergence than FedAvg. FedProx, starting with the highest initial accuracy (0.9683), maintained consistent progress and reached 0.9989 by the final round, confirming its robustness and effectiveness in handling heterogeneous and non-IID data distributions.

These findings validate the study's objective by demonstrating that federated optimization algorithms can effectively manage data heterogeneity and ensure secure, distributed model training for sensitive educational analytics. However, the experiments were conducted in a simulated federated environment involving a limited number of institutions, which may not fully represent large-scale or real-world deployments.

Future work will address these limitations by incorporating a broader range of educational institutions, exploring asynchronous and cross-silo FL configurations, and integrating additional behavioral and contextual variables to enhance predictive accuracy, adaptability, and robustness. Previous studies have explored centralized machine learning for mental health analysis and early intervention [31]; however, these relied on centralized data aggregation, leading to privacy and scalability challenges.

This research advances the field by introducing a federated learning-based approach that ensures data confidentiality while

maintaining model accuracy and adaptability across heterogeneous educational environments. Overall, it establishes a foundation for ethical, data-driven, and privacy-aware educational analytics in Kazakhstan and beyond, supporting early intervention strategies and strengthening national efforts to promote student mental health and academic resilience.

REFERENCES

[1] G. P. Reddy and Y. V. Pavan Kumar, "A beginner's guide to federated learning," in 2023 Intelligent Methods, Systems, and Applications (IMSA), 2023, pp. 557–562. doi:10.1109/IMSA58542.2023.10217383.

[2] J. -W. Lin, T. Kim, H. -J. Tu and P. -H. Kuo, "Adaptive Model Transfers and Aggregations for Efficient Federated Learning in IoT Edge Systems with Non-IID Data," 2023 6th International Conference on Data Science and Information Technology (DSIT), Shanghai, China, 2023, pp. 56-60, doi: 10.1109/DSIT60026.2023.00017.

[3] S. M. Jalal, M. R. Hasan, M. A. Haque and M. G. R. Alam, "A Horizontal Federated Random Forest for Heart Disease Detection from Decentralized Local Data," 2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC), Hyderabad, India, 2022, pp. 191-196, doi: 10.1109/R10-HTC54060.2022.9929490.

[4] J. Mills, J. Hu and G. Min, "Multi-Task Federated Learning for Personalised Deep Neural Networks in Edge Computing," in IEEE Transactions on Parallel and Distributed Systems, vol. 33, no. 3, pp. 630-641, 1 March 2022, doi: 10.1109/TPDS.2021.3098467.

[5] Pranto, M.A., & Al Asad, N. (2021). A Comprehensive Model to Monitor Mental Health based on Federated Learning and Deep Learning. 2021 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON), 18-21.

[6] J. -W. Lin, T. Kim, H. -J. Tu and P. -H. Kuo, "Adaptive Model Transfers and Aggregations for Efficient Federated Learning in IoT Edge Systems with Non-IID Data," 2023 6th International Conference on Data Science and Information Technology (DSIT), Shanghai, China, 2023, pp. 56-60, doi: 10.1109/DSIT60026.2023.00017.

[7] Q. Ma et al., "FedSA: A Semi-Asynchronous Federated Learning Mechanism in Heterogeneous Edge Computing," in IEEE Journal on Selected Areas in Communications, vol. 39, no. 12, pp. 3654-3672, Dec. 2021, doi: 10.1109/JSAC.2021.3118435.

[8] Z. Jiang et al., "Computation and Communication Efficient Federated Learning With Adaptive Model Pruning," in IEEE Transactions on Mobile Computing, vol. 23, no. 3, pp. 2003-2021, March 2024, doi: 10.1109/TMC.2023.3247798.

[9] T. Vu et al., "User selection approaches to mitigate the straggler effect for federated learning on cell-free massive MIMO networks," arXiv preprint arXiv:2009.02031, 2020.

[10] J. Le et al., "Federated Continuous Learning With Broad Network Architecture," in IEEE Transactions on Cybernetics, vol. 51, no. 8, pp. 3874-3888, Aug. 2021, doi: 10.1109/TCYB.2021.3090260.

[11] J. P. Albrecht, "How the GDPR will change the world," Eur. Data Prot. Law Rev., vol. 2, pp. 287–289, 2016. doi: 10.1093/IDPL/IPX023.

[12] X. Zhang, Q. Wang, Z. Ye, H. Ying and D. Yu, "Federated Representation Learning With Data Heterogeneity for Human Mobility Prediction," in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 6, pp. 6111-6122, June 2023, doi: 10.1109/TITS.2023.3252029.

[13] K. Bonawitz et al., "Practical secure aggregation for privacy-preserving machine learning," in Proc. 2017 ACM SIGSAC Conf. Comput. Commun. Secur., 2017 doi:10.1145/3133956.3133982.

[14] J. Konecný et al., "Federated learning: Strategies for improving communication efficiency," arXiv preprint arXiv:1610.05492, 2016.

[15] O. R. A. Almanifi et al., "Communication and computation efficiency in federated learning: A survey," Internet Things, vol. 22, p. 100742, 2023. doi: 10.1016/j.iot.2023.100742.

[16] H. Zhang, L. Liu, and L. Lu, "Federated scientific machine learning for approximating functions and solving differential equations with data heterogeneity," arXiv preprint arXiv:2410.13141, 2024. doi: 10.48550/arXiv.2410.13141.

[17] T. Sun, D. Li, and B. Wang, "Decentralized federated averaging," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, pp. 4289–4301, 2021. doi: 10.1109/TPAMI.2022.3196503.

[18] S. J. Reddi et al., "Adaptive federated optimization," arXiv preprint arXiv:2003.00295, 2020.

[19] M. A. Hidayat, Y. Nakamura and Y. Arakawa, "Privacy-Preserving Federated Learning With Resource-Adaptive Compression for Edge Devices," in IEEE Internet of Things Journal, vol. 11, no. 8, pp. 13180-13198, 15 April15, 2024, doi: 10.1109/JIOT.2023.3347552.

[20] H. Ko et al., "Joint Client Selection and Bandwidth Allocation Algorithm for Federated Learning," in IEEE Transactions on Mobile Computing, vol. 22, no. 6, pp. 3380-3390, 1 June 2023, doi: 10.1109/TMC.2021.3136611.

[21] C. Wu, Z. Li, F. Wang and C. Wu, "Learning Cautiously in Federated Learning with Noisy and Heterogeneous Clients," 2023 IEEE International Conference on Multimedia and Expo (ICME), Brisbane, Australia, 2023, pp. 660-665, doi: 10.1109/ICME55011.2023.00119.

[22] M. Favero, T. Marchioro and C. Schiavo, "Federated Forests With Differential Privacy for Distributed Wearable Sensors," 2024 IEEE International Conference on Aerospace Electronics and Remote Sensing Technology (ICARES), Yogyakarta, Indonesia, 2024, pp. 1-6, doi: 10.1109/ICARES64249.2024.10767941.

[23] X. Feng, Q. Shen, C. Li, Y. Fang and Z. Wu, "Privacy Preserving Federated Learning from Multi-Input Functional Proxy Re-Encryption," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 6955-6959, doi: 10.1109/ICASSP48485.2024.10446283.

[24] L. Liu and D. Zhou, "Analysis of regularized federated learning," Neurocomputing, vol. 611, p. 128579, 2024. doi: 10.1016/j.neucom.2024.128579.

[25] Koh, Y., Lee, C., Ku, Y., & Lee, U. (2023). Data Visualization for Mental Health Monitoring in Smart Home Environment: A Case Study. 2023 Workshop on Visual Analytics in Healthcare (VAHC), 53-55.

[26] Wang, P. (2024). Analysis of Adolescent Mental Health Based on Machine Learning Algorithm. 2024 3rd International Conference for Advancement in Technology (ICONAT), 1-5.

[27] L. Shi et al., "Data poisoning attacks on federated learning by using adversarial samples," 2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI), Shijiazhuang, China, 2022, pp. 158-162, doi: 10.1109/ICCEAI55464.2022.00041.

[28] W. Qiu, W. Ai, H. Chen, Q. Feng and G. Tang, "Decentralized Federated Learning for Industrial IoT With Deep Echo State Networks," in IEEE Transactions on Industrial Informatics, vol. 19, no. 4, pp. 5849-5857, April 2023, doi: 10.1109/TII.2022.3194627.

[29] Y. Deng and X. Yan, "Federated Learning on Heterogeneous Opportunistic Networks," 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Nanjing, China, 2024, pp. 447-451, doi: 10.1109/AINIT61980.2024.10581819.

[30] Tang and T. -H. Chang, "FedLion: Faster Adaptive Federated Optimization with Fewer Communication," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 13316-13320, doi: 10.1109/ICASSP48485.2024.10447045.

[31] Kataru, S., King, K., & Fernando, L. (2024). Machine Learning-Based Early Detection and Intervention for Mental Health Issues in Children. 2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC), 2001-2007.

Exploring Hallucination in Large Language Models

A Comparative Study of GPT-4o and GPT-4o-mini in Medical Domains

Nesreen M. Alharbi[1], Thoria Alghamdi[2], Raghda M. Alqurashi[3] ,Reem Alwashmi[4], Amal Babour[5]*, Entisar Alkayal[6]
Department of Computer Science-Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Rabigh 21911, Saudi Arabia[1]
THE LIGHT-Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Rabigh 21911, Saudi Arabia[1, 6]
Department of Information Systems-Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia[2, 5]
Department of Computer Science-Jamoum University College, Umm Al-Qura University, Al Jumum 25371, Saudi Arabia[3]
Department of Information Technology-Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Rabigh 21911, Saudi Arabi[4]

Abstract*—Large Language Models such as GPT-4o and GPT-4o-mini have shown significant promise in various fields. However, hallucination, when models generate inaccurate information, remains a critical challenge, especially in domains that require high accuracy, such as the healthcare field. This study investigates hallucinations in two different LLMs, focusing on the healthcare domain. Four different experiments were defined to examine the two models' memorization and reasoning abilities. For each experiment, a dataset with 193,155 multiple-choice medical questions from postgraduate medical programs was prepared by splitting it into 21 subsets according to medical topics. Each subset has two versions: one with the correct answers included and one without them. Accuracy and compliance were evaluated for each model. Models' adherence to requirements in prompts was assessed. Also, the correlation between size and accuracy was tested. The experiments were repeated to evaluate the models' stability. Finally, the models' reasoning was evaluated by human experts who assessed the models' explanations for correct answers. The results revealed poor rates of accuracy and compliance for the two models, with rates below 70% and 75%, respectively, in most datasets; yet, both models showed low uncertainty (3%) in their responses. The findings showed that the accuracy was not affected by the size of the dataset provided to the models. Also, the results indicated that GPT-4o-mini demonstrates greater performance stability compared to GPT-4o. Furthermore, the two models provided acceptable justifications for choosing the correct answer in most cases, according to 68.8% of expert questionnaire participants who agreed with both models' justifications. According to these results, both models cannot be relied upon when accuracy is critical, even though GPT-4o-mini slightly outperformed GPT-4o in providing the correct answers. The findings highlight the importance of improving LLM accuracy and reasoning to ensure reliability in critical fields like healthcare.

Keywords*—ChatGPT; GPT-4o; GPT-4o-mini; hallucination; healthcare; large language models

I. Introduction

One development in artificial intelligence (AI) applications is large language models (LLMs), which include several technologies such as deep learning and natural language processing (NLP), reinforcement learning, and transformers to produce new content from available material. Today, there are many LLMs that have been developed by different vendors. These models follow the same process and start with training the LLM on a large amount of data from the Internet to produce new content. However, these models differ in their training methods and the technologies used to generate and predict output data. Most of these models rely on language general context, and linguistic rules to predict subsequent words based on the preceding ones using technology [1].

LLM models have been applied in many fields to generate content ranging from answering questions, executing commands and instructions, summarizing, translating, and conducting long conversations with users [2]. One well-known company in this field is OpenAI. In November 2022, it made its LLM, ChatGPT, available to the public. Following that, OpenAI launched several versions of ChatGPT. For instance, in March 2023, GPT-4 was introduced, followed by GPT-4o in May 2024. These new releases had a high impact on NLP technology. The accuracy of understanding the context of input data and generating the response has enhanced in comparison to the old models. In addition, GPT-4o shows strong performance across many benchmarks [3], and supports multi-modal content generation.

LLMs have been widely used in several fields, including education, healthcare, industry, business, and marketing, to enhance productivity and quality as well as reducing effort and saving time [1], [4], and [5]. However, the use of LLMs may involve some ethical issues such as bias and hallucinations which can affect data accuracy and the model reliability [6], [7], and [8]. Hallucinations may occur when the models been trained with invalid and insufficient data, thus LLM generates inaccurate or false outputs [9], and [10]. Several studies have identified factors to minimize the possibility of hallucination. These factors include improving the quality of training data, developing methods to detect and correct errors, and enhancing the models' ability to think and reason to distinguish accurate information from false [10].

The problem of hallucination is that it looks like a piece of meaningful and valuable information, but, in fact, it is inaccurate and invalid. Generally, hallucinations can occur for several reasons, including a lack of adequate information, which leads to inaccurate results during analysis. In addition, poor data

*Corresponding author.

quality will generate models that rely on low-quality data, reducing result accuracy and increasing the likelihood of hallucinations [9]. Moreover, some instructions or questions given to the LLM models may be unclear or can result in multiple interpretations according to different contexts. Furthermore, most LLM models lack reasoning, critical thinking, and analytical abilities, although companies have focused on improving these issues in the current updated versions [10].

This research contributes by examining two models of ChatGPT: GPT-4o and GPT-4o-mini and comparing the hallucination of the two models in the healthcare field. The motivation behind this work comes from the hallucination issue in LLMs, which is particularly critical in fields such as healthcare, where accurate responses are essential and any misinformation could have severe or even fatal consequences. Reducing hallucination in LLMs is essential for ensuring their reliability in critical fields. Thus, the objective of this research is to evaluate the two models' responses to medical questions. Different sets of medical exam questions in different medical topics for postgraduate medical programs were prepared by including or excluding the correct answers to test the models' ability to reason and distinguish the absence of correct answers. Experiments were designed to evaluate GPT-4o and GPT-4o-mini models on memorization and reasoning, where accuracy, compliance, uncertainty, and reasoning were measured for each experiment. Also, this research examines the relationship between the size of the dataset and the models' accuracy, as well as the impact of repeating the test on the models' accuracy.

The rest of the research is organized as follows: Section II reviews the literature. In Section III, the methodology is addressed. A detailed explanation of the different experiments and their results is presented in Section IV and Section V, respectively. Section VI discusses the findings of the study. Concluding remarks and a brief of future works are given in Section VII.

II. Literature Review

Models of large languages like LLama 3 [1], and GPT-4o [11] have the capability to generate loads of information as needed in several domains, including professional and crucial [12], [4]. However, in some cases, hallucination phenomena may occur in LLMs [2] when providing information, which can vary, ranging from minor fabrications to major misconceptions.

Regardless of the effectiveness of the prompting techniques, researchers have found that LLMs can produce erroneous responses that mimic actual statements but contain unsubstantiated information [13]. Others have confirmed that when hallucinations occur in LLMs, the resulting errors can negatively impact the user experience, leading to confusion, distrust, and decreased enthusiasm for using these AI models [7] and [14]. Ji et al. [7] define hallucinations in LLMs as the natural generation of information; however, they may lack meaning or be inconsistent with the actual content of the source. Zhang et al. [15] have standardized the definition of hallucinations in LLMs into three categories: input-conflicting hallucinations, context-conflicting hallucinations, and fact-conflicting hallucinations. With input-conflicting hallucinations, the content that LLMs produce conflicts with what the user has input. With context-conflicting hallucinations, the content produced varies between two different attempts. Lastly, fact-conflicting hallucinations occur when the information produced contradicts fundamental knowledge. Several factors can significantly cause hallucination in LLMs, including inadequate data samples in the training dataset or utilizing algorithms with extremely uncertain sampling [7] and [16].

Therefore, dependence on language generative technologies, where hallucinations in LLMs may occur, raises serious concerns. Given these challenges, it is particularly concerning when LLMs are applied, particularly in critical domains like healthcare. This is due to the significant impact on people's lives [17], including incorrect clinical decision-making and delayed or improper treatment [8]. It is possible to train LLMs to perform remarkably well on a range of medical and healthcare tasks [11], [4], pass medical examinations [5], [18], [19], and generate relevant texts about medical topics. However, in a question-answering task, LLMs may provide an answer that appears reasonable but is factually inaccurate, while in content-generation tasks, the model may also generate cohesive narratives or explanations based on fictitious facts or events in content generation tasks. For example, in response to the question, "What are the common side effects of metformin?" an LLM might say, "Medication side effects include nausea as well as trouble breathing", which is slightly incorrect [20]. The reliability of LLMs is a major concern for both healthcare organizations and patients, especially if the model hallucinates and produces incorrect answers to medical questions. Any incorrect information can significantly impact the patient's health and the healthcare organization's reputation. Thus, hallucinations in LLMs might be extremely challenging in the medical domain [21], [22], where integrity and dependability are crucial [23] to maintain both patient satisfaction and trustworthiness while receiving healthcare.

There are several studies that highlighted the performance of different LLM models, including MCQ examination in the domain of health care [24, 25, 26, 27]. Each research has evaluated the performance of hallucinations based on one or more LLM models like ChatGPT3, ChatGPT 4 [24, 25, 26] while [25] compares ChatGPT-4 with three other different models, namely ChatGPT, QWen 2.1, and Ernie 4.0. Nevertheless, [27] has examined the performance of ChatGPT solely. In addition, the dataset language used in each research varies between English and Non-English, aimed to measure both the model's accuracy and reasoning.

Moreover, [24] focused on evaluating ChatGPT-4 performance on the Japanese medical licensing examination dataset. Four aspects of the model were examined: accuracy, category sub-score, effect of translation, and error analysis. The model scored 82.7% and 77.2% for the essential questions and the basic and clinical questions, respectively. On the other hand, [25] has tested the performance of the GPT-4.0, ChatGPT, QWen 2.1, and Ernie 4.0 models on a Chinese dataset that has been evaluated by experts; the models' accuracy, performance, and reasoning were measured. The research measures model reasoning by calculating the rates of fact hallucinating, fact fabrication, Instruction Inconsistency, and logical inconsistency. For all models, Ernie 4.0 achieved better performance compared to GPT-4.0.

Furthermore, [26] compared the performance of ChatGPT-4 in answering medical MCQ questions with the answers coming from medical students. This study showed that ChatGPT-4 accuracy outperformed the students with 73.7%, 66.7% respectively. On the other hand, 78% of the students achieved at least 90% accuracy, which is higher than the accuracy rate of ChatGPT-4. Finally, in [27], ChatGPT's accuracy and reasoning were evaluated using MCQ datasets from the United States Medical Licensing Examination (Steps 1 and 2), using two different datasets with and without a provided hint to the model. Having four different versions of the question sets, the study revealed accuracy over datasets for the first dataset (Step1) exam as 44%, (Step2) as 42% and the second dataset (Step1) as 64.4%, and (Step2) as 57.8%.

Although these studies tested the abilities of different LLMs, none of them evaluated the ability of the model while excluding the correct answers. Also, none of these studies have tested hallucination on Chat-GPT-4mini and measured the correlation between the model's accuracy and the size of the dataset, nor tested the model's stability while answering medical questions.

III. Materials and Methods

The dataset for this research [28] consists of multiple-choice questions (MCQs) with fine-grained human-labeled classes in different medical fields at the graduate level, designed as a prerequisite for admission to postgraduate medical programs. These questions are designed to assess the skills of medical professionals. The dataset samples contain questions with answer options, correct answers, and explanations of the solutions. To ensure that all questions could be answered by text input, a few steps were taken to clean the data. First, the questions and options were proofread. In addition, to improve the quality of the dataset, the content was supervised by humans. Some tools were used to prepare the data, e.g., a spellchecker to identify and correct some cases, such as extra white spaces or missing options. Any question with an inconsistent format or questions without the correct answer option was removed. In addition, questions without a null candidate or those requiring external information were deleted. Questions containing specific keywords, such as "equation" or "India", were removed. In addition, all duplicate questions were removed. The final version of the dataset contained 193,155 questions.

To enable the generalization and re-usability of the models, the dataset was split by exams rather than by the questions given. The training set consisted of collected mock and online test series, while the test set comprised MCQs from the AIIMS PG exams (years 1991–the date at the time of when you compiled the dataset). The NEET PG exam MCQs (years 2001–the date at the time of when you compiled the dataset) were used to make the development set, aiming to approximate a real exam evaluation. The dataset consisted of 183,000 training examples, 4000 in the test set, and 6000 in the development set. To avoid overlapping between questions from the training, test, and development datasets, it was ensured that the test and development sets contained questions distinct from the training data. By calculating the Levenshtein distance between each pair of questions in the dataset, a question was excluded from the development and test sets if its similarity to any other question exceeded 0.9. For this research, the dataset was split into 21 smaller subsets, each corresponding to specific topics covered by the questions. These topics were Surgery, Social and Preventive, Skin, Radiology, Psychiatry, Physiology, Pharmacology, Pediatrics, Pathology, Orthopedics, Ophthalmology, Microbiology, Medicine, Gynecology and Obstetrics, Forensic Medicine, ENT, Dental, Biochemistry, Anatomy, Anesthesia, and Unknown. The Unknown dataset included all the questions that could not be categorized into any of the other 20 topics. After splitting, the dataset was divided into 21 sub-datasets, and 10% of the questions from each sub-dataset were randomly selected. This reduced the number of questions used to evaluate the targeted models to 100,000. Then, to prepare the data, all unnecessary records were dropped, and only the question, the four answer options, and the correct answer were retained. These datasets were used to examine the targeted models in the set of experiments where the correct answer is provided to the model with the set of possible answers. The final structure of the questions in this dataset is shown in Fig. 1.

```
Question: {Question}
Opt0: {Option1}
Opt1: {Option2}
Opt2: {Option3}
Opt3: {Option4}
Answer: {Correct Choice (0-3)}
Opt0, 1, 2, or 3 includes a correct choice
```

Fig. 1. Structure of question set with correct answer provided.

Then, new datasets were generated from the previous datasets to conduct another set of experiments where the correct answer was not included in the possible choices. Instead, the correct choice was replaced by (NONE). The final structure of the questions in these datasets is shown in Fig. 2.

```
Question: {Question}
Opt0: {Option1}
Opt1: {Option2}
Opt2: {Option3}
Opt3: {Option4}
Answer: {Correct Choice (0-3)}
Opt0, 1, 2, or 3 includes a NON choice, no
    correct answewr provided.
```

Fig. 2. Structure of question set with NO correct answer provided.

IV. Experiments

In order to test hallucination in LLMs, two different models were tested, focusing on two key capabilities: memorization and reasoning. This study examined two well-known models: GPT-4o and GPT-4o-mini. The experiments conducted in this research were implemented using Google Collab, with Python 3 and a CPU hardware accelerator. Four separate experiments were designed to evaluate the models: two for GPT-4o and two for GPT-4o-mini. Each model was tested for both memorization and reasoning abilities. In addition, two of the experiments were repeated to examine the ability of the two models to improve their levels of memorization and reasoning.

A. GPT-4O Hallucination Evaluation

1) Analyzing Memorization and Reasoning with Correct Answer (MWRCA)

The objective of this experiment was to evaluate the hallucination of the GPT-4o model at two different levels:

memorization and reasoning. For this purpose, a benchmark consisting of multiple-choice questions, each with four answer options, was used that covered 21 different domains (see Section II). Among the four options for each question, one was the correct answer. The evaluation method (GPT-4o: MWRCA) began by feeding the model this set of questions, then asking it to select the correct answer and provide a reason for its selection. Additionally, the model was asked to explain why the other options were incorrect. After collecting the model's outputs, its memorization ability was evaluated based on the number of correct answers it chose, while its reasoning ability was assessed by analyzing the justifications the model provided for selecting the correct answer and avoiding the incorrect ones. Fig. 3 shows the prompt that asked the model to comply with a specific format to generate the output response. In this test, the model was allowed to respond "(IDK)" if it was uncertain of the correct answer. However, the expectation in the GPT-4o: MWRCA experiment was that the model would choose the correct answer rather than selecting an incorrect option or responding with "IDK". The results produced by the model were then saved in a CSV file.

```
You are a knowledgeable and meticulous specialist in the medical field.
Your role is to analyze multiple-choice questions along with their
provided options. Given a multiple-choice question with four possi-
ble answers (0, 1, 2, 3), determine the correct answer and ‘ deliver a
comprehensive explanation for its accuracy.
Additionally, combine the explanations for why the remaining options are
incorrect into a single column labeled “Incorrect Explanations”.
The output must strictly follow this structure:
- Column 1: Question Header (the question text or ID).
- Column 2: Correct Option Index (e.g., 9, 1, 2, 3, or IDK or IDK).
- Column 3: Correct Explanation (a single, clear sentence ezolaning the
correct answer).
- Column 4: Incorrect Explanations (all incorrect option explanations
combined lnto one clear sentence, separated by semicolons).
Ensure that:
 1. Each column is clearly separated by a single comma (…).
 2. No explanation spills over into another column. Keep explanations and
limited to their respective columns.
 3. If uncertaln about the correct answer, write ‘IDK’ in Column 2, leave
Column 3 blank, and provide reasoning in Column 4
Provide only one row of output in the format:“Question ‘Expalanations”
Strict adherencce to this format is required  Any deviation is unacetable.
```

Fig. 3. Prompt instruction to ChatGPT-4o.

To calculate the accuracy of the model, its output answers were compared to the actual correct answers. For each correct response, the model received a score of one. The total score was then divided by the number of questions in each topic, as shown in Formula (1):

$$\text{Accuracy} = \frac{\text{TotalNumberofCorrectAnswers}}{\text{TotalNumberofQuestions}} \times 100 \quad (1)$$

To calculate the compliance of the model, the model's output was compared to the correct structure of the output format (question, correct option index, correct explanation, incorrect explanation). The model received a score of one if it followed the specific format; otherwise, it received a score of zero. The total score of the model was then divided by the total number of questions in each topic, as shown in Formula (2):

$$\text{Compliance} = \frac{\text{ScoreofCompliance}}{\text{TotalNumberofQuestions}} \times 100 \quad (2)$$

Finally, the uncertainty of the model was calculated based on the number of IDK responses it provided. The total number of IDK responses was then divided by the total number of questions to calculate the percentage of uncertainty, as shown in Formula (3):

$$\text{Uncertainty} = \frac{\text{TotalNmuberofIDKresponses}}{\text{TotalNumberofQuestions}} \times 100 \quad (3)$$

2) Analyzing Memorization and Reasoning with No Correct Answer (MWRNCA)

The GPT-4o: MWRNCA experiment followed the same structure as the GPT-4o: MWRNCA experiment in terms of constructing the prompt. The key difference between the two experiments is the structure of the benchmark used to evaluate the models. In this experiment, the set of questions presented to the model had no correct answer in the available choices for each question. Instead, options such as "Blank", "NONE", or "NONE OF THE ABOVE" were included to represent "No Correct Answer Provided". Fig. 4 shows an example of a question from the dataset. The expectation for the results was that the model would choose "Blank", "NONE", or "NONE OF THE ABOVE" as an answer for each question. To run GPT-4o: MWRNCA, the model was not given any hints about removing the correct answers from the dataset. The model was expected to know this by itself.

```
1 Question: {Post operative muscle ache is
      caused by}
2 Opt1: {d-TC}
3 Opt2: {None}
4 Opt3: { Gallamine}
5 Opt4: { Pancuronium}
```

Fig. 4. Question sample from the anaesthesia dataset.

The GPT-4o responses were evaluated based on accuracy, compliance, and uncertainty. Accuracy was calculated based on the number of "NONE" responses provided by the model, relative to the total number of questions [see Formula (4)]. Further, the percentage of the model compliance and uncertainty was calculated following Formula (2) and (3), respectively.

$$\text{Accuracy} = \frac{\text{TotalNmuberofNONEresponses}}{\text{TotalNumberofQuestions}} \times 100 \quad (4)$$

B. GPT-4O-Mini Hallucination Evaluation

1) Analyzing Memorization and Reasoning with Correct Answer (MWRCA)

In this experiment, GPT-4o-mini hallucination in memorization and reasoning was evaluated. Following the methodology of the GPT-4o: MWRCA experiment, questions from 21 different medical topics were passed to the model to answer. The same prompt defined in Section II was used with GPT-4o-mini. The model's responses were evaluated based on the model accuracy, compliance, and uncertainty, following Formula (1), (2), and (3), respectively.

2) Analyzing Memorization and Reasoning with No Correct Answer (MWRNCA)

In this experiment, the hallucination behavior of GPT-4o-mini was evaluated at both the memorization and reasoning levels. The prompt provided to the model was similar to the one in Section I, with the only change the model's name, i.e., GPT-4o-mini instead of GPT-4o. The dataset used in this experiment was the one that included no correct answers in the provided

answer choices. The model was expected to respond with "NONE"; any other response reduces its accuracy according to Formula (4). Compliance and uncertainty percentages are calculated using Formula (2) and (3), respectively.

V. Results

In each of the four experiments, the compliance rate, accuracy, and uncertainty were measured for each dataset. Additionally, the correlation between the size of the dataset and the accuracy was analyzed, as well as the learning rate after repeating the experiments. Lastly, the models were prompted to provide reasoning for their answers, and these justifications were sent to medical experts for evaluation of the models' reasoning.

A. Compliance Rate

The compliance rate is defined as the number of compliant responses divided by the total number of responses. As shown in Fig. 5, the compliance rate of both models slightly decreased across all datasets when the correct answer was replaced with "NONE" in the given options; however, this drop is insignificant. GPT-4o-mini outperformed GPT-4o in terms of compliance in nearly all experiments. The only exception was observed on the ENT dataset, where GPT-4o complied with the requested response format on about 99.8% of the questions when the correct answer was given in the options. On average, the compliance rates for the GPT-4o: MWRCA, GPT-4o: MWRNCA, GPT-4o-mini: MWRCA, and GPT-4o-mini: MWRNCA were 60.98%, 57.92%, 76.30%, and 72.94%, respectively.

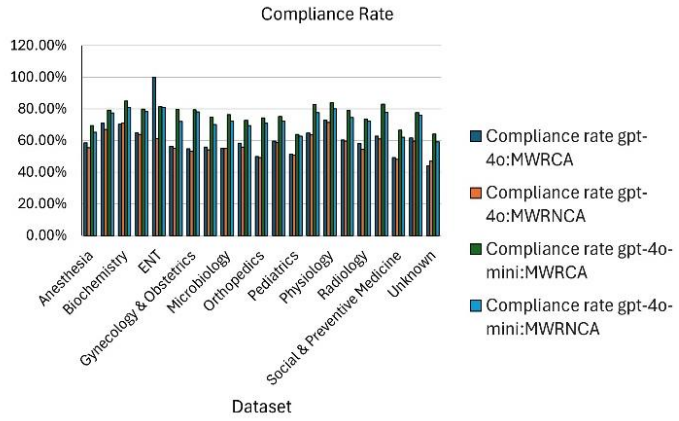


Fig. 5. The compliance rate of each experiment on each dataset.

B. Accuracy

The accuracy of each experiment across all datasets is shown in Fig. 6. Although GPT-4o-mini slightly outperforms GPT-4o on most datasets, both models perform poorly with accuracy rates below 70% on all datasets, except the ENT dataset, where GPT-4o achieves an accuracy of 81%. Both models show a significant drop in performance when the correct answer is replaced with "NONE", with average decreases of 75% for GPT-4o and 66% for GPT-4o-mini. The average accuracy for GPT-4o: MWRNCA, GPT-4o- mini: MWRCA, and GPT-4o-mini: MWRNCA are 48.60%, 11.90%, 53.13%, and 17.38%, respectively. These results indicate that both models are highly hallucinatory when applied to the given medical datasets.

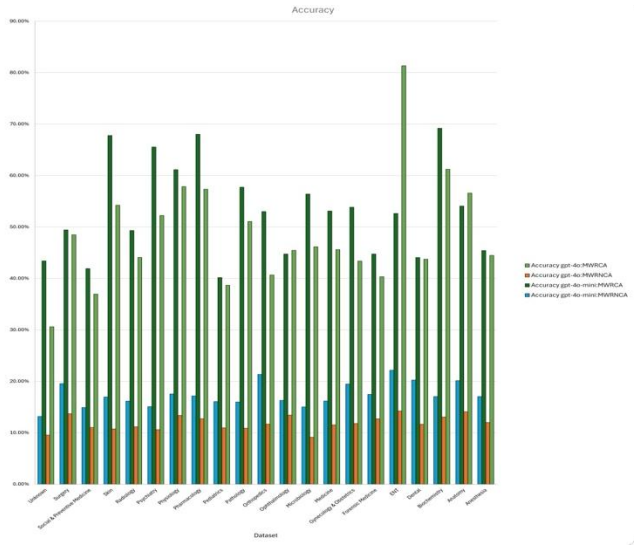


Fig. 6. The accuracy of each experiment on each dataset.

C. Uncertainty

When the correct answer is listed within the options, the models had less than 3% uncertainty on all datasets, as seen in Fig. 7. Additionally, GPT-4o reported a higher uncertainty than GPT-04-mini on all datasets. Notably, despite the low accuracy of all models as mentioned above, they all reported low uncertainties, averaging 0.88%, 26.63%, 0.59%, and 13.15% for GPT-4o: MWRCA, GPT-4o: MWRNCA, GPT-4o-mini: MWRCA, and GPT-4o-mini: MWRNCA, respectively.

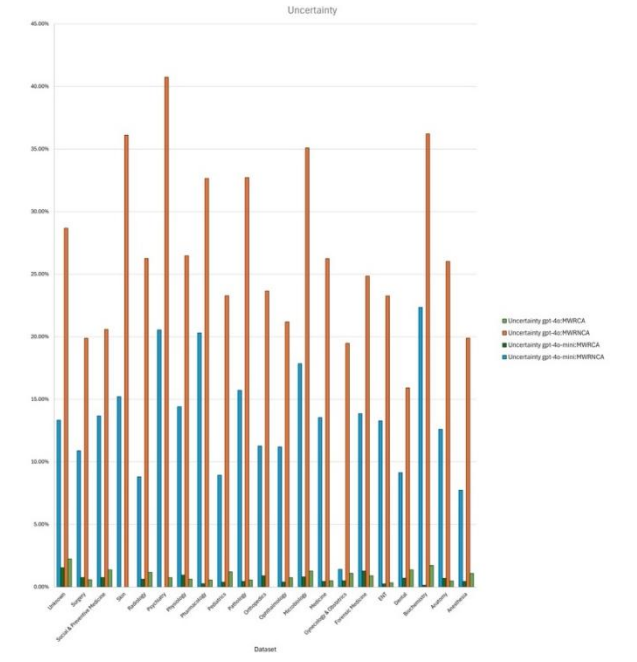


Fig. 7. The uncertainty of each experiment on each dataset.

D. Correlation

The correlation factors between the experiments' accuracy and the dataset sizes are all below 0.05, as seen in Fig. 8. The results indicate no clear correlation between dataset size and model performance.

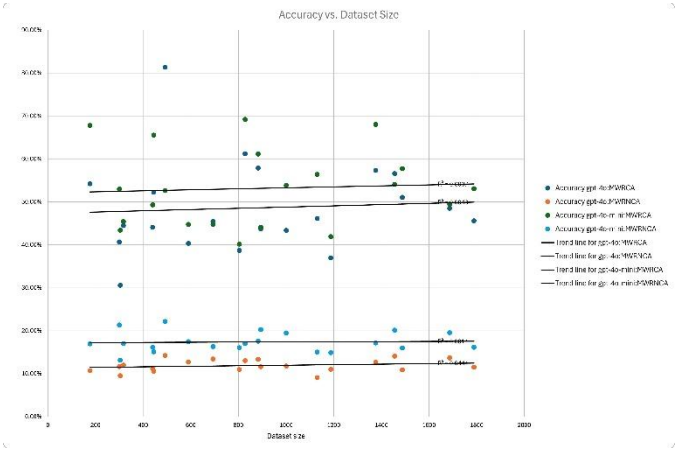


Fig. 8. The correlation between the dataset size and the accuracy.

E. Model Stability

To measure the models' performance stability, each of the four experiments was rerun on the following 10 selected datasets: Pathology, Pediatrics, Pharmacology, Physiology, Psychiatry, Radiology, Skin, Social and Preventive Medicine, Surgery, and Unknown. The computed difference between the accuracies of the two iterations of each of the four experiments, as in Formula (5), was 30.74% for GPT-4o with the correct answer, 1.75% for GPT-4o-mini with the correct answer, 3.72% for GPT-4o with "NONE" instead of the correct answer, and 1.64% GPT-4o-mini with "NONE'" replacing the correct answer. These results indicate that GPT-4o-mini demonstrates greater performance stability compared to GPT-4o.

$$\text{ChangeInAccuracy} = \text{SecondRunAccuracy} - \text{FirstRunAccuracy} \quad (5)$$

F. Reasoning

An electronic questionnaire was completed by 22 medical doctors. The questionnaire consisted of six sections, each representing a different medical field. Each section contained randomly selected case-based questions that both models answered correctly. Along with each question were two justifications, one from GPT-4o and the other from GPT-4o mini, along with the model's selected answer and its justification.

Participants were asked to evaluate the output of each model by selecting one of four possible options for each question which are:

- Agree with both GPT-4o and GPT-4o mini justifications.
- Agree with GPT-4o justification but disagree with GPT-4o mini justification.
- Disagree with both GPT-4o and GPT-4o mini justifications.
- Disagree with GPT-4o justification but agree with GPT-4o mini justification.

The results of this questionnaire revealed that the first choice had the highest percentage (68.8%) among all options. The third-choice percentage was the lowest (3.3%). The fourth option received 15.5%, while the second option received 12.4% of participants' selections. These results indicate that the justifications provided by both GPT-4o and GPT-4o-mini were considered reasonable for most of the questions.

Fig. 9 summarizes the average compliance, accuracy, and uncertainty for each experiment across all datasets. While GPT-4o-mini slightly outperformed GPT-4o in answering multiple-choice medical questions, both models frequently hallucinated and were unreliable for medical decision-making. Hallucination rates increased further when the correct answer was replaced with "NONE" as an option. Despite low overall accuracy, the models reported very low uncertainty when the correct answer was explicitly provided among the choices. However, substituting the correct answer with "NONE" reduced accuracy while increasing reported uncertainty. In terms of correlation, no relationship was found between dataset size and model accuracy, suggesting that repeated exposure to a topic does not improve performance. Additionally, retesting revealed that GPT-4o-mini's results remained stable across experiments, whereas GPT-4o's performance varied in the experiment with the correct answers explicitly included.

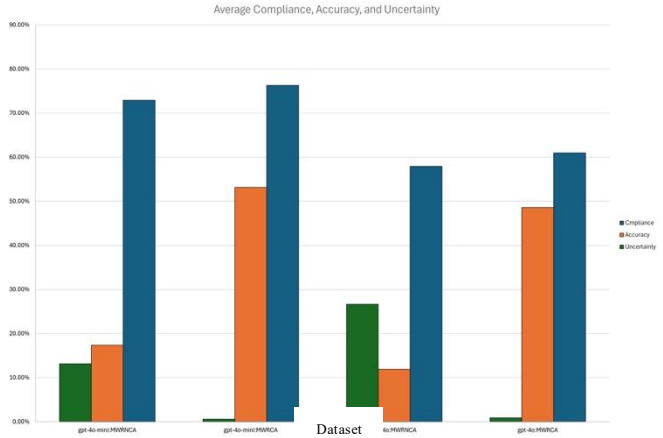


Fig. 9. The overall average of compliance, accuracy, and uncertainty.

VI. Discussion

The results of this study highlight significant findings regarding the occurrence of hallucinations in two large language models, GPT-4o and GPT-4o-mini, in the context of healthcare-related medical questions. Both models demonstrated notable challenges in their ability to reliably generate correct answers, with accuracy rates falling below 70% in most datasets. Although GPT-4o-mini performed better than GPT-4o in terms of compliance and accuracy, it still exhibited considerable hallucinatory behavior and weak adherence to a given task instruction. On the other hand, the accuracy of the two models dropped when the correct answer was replaced with a "NONE" option. This drop indicates that both models could not recognize the absence of the correct answer and generate unreliable answers without real understanding which highlights the limitations of these LLMs when response correctness is a key factor. Moreover, the high level of compliance combined with the drop in accuracy suggests that these LLMs cannot be relied upon, especially if used in high-stakes contexts such as healthcare. Furthermore, the study's experiments demonstrated a lack of correlation between the size of the question set sent to the model and its performance, showing no relation between

changing the size of the data and the model's accuracy which indicated that model hallucination is not sensitive to a given task size. Finally, re-running the experiments on smaller and specific datasets showed that GPT-4o-mini has minimal accuracy differences between runs compared to GPT-4o. This indicates that GPT-4o-mini may be a more stable model in real-world applications, although it remains at risk of hallucinations.

Comparing the results of this study with previous studies conducted on LLMs hallucination in the medical field, the presented results are consistent with the previous studies [24, 25, 26, 27] that asserted that the LLMs, including different GPT-versions exhibit hallucination and show low accuracy. In the context of evaluating model hallucination in different languages, two studies [24,25] conducted in non-English languages showed similar results to the studies conducted in English [26,27] and to this study, indicating that LLMs exhibit hallucination despite the language used. However, this research is distinguished from the previous studies in the variety of evaluation criteria and the medical topics. In this study, multiple tests were designed to measure specific aspects of the models, providing deeper insight into their performance. In addition to recall and reasoning, the study assessed models' stability and their correlation with the size of the question set. Also, the study explores the models' behavior in the absence of correct answers. Whereas previous studies tested the medical dataset ranged from one to four topics, this study's experiments were conducted on 21 medical topics to provide a deeper insight into hallucination and to avoid the limitation of a broad analysis.

In conclusion, this study concluded that GPT-4o and GPT-4o-mini exhibit hallucination in the medical field, highlighting the need for further study to test different LLMs and apply on different datasets. Finally, this research recommends improving the LLMs and using them cautiously, considering the hallucination in their responses.

VII. Conclusion and Future Work

The main findings of this study present the results of the compliance, accuracy, and uncertainty of each experiment across all datasets. Although GPT-4o-mini slightly outperformed GPT-4o in correctly answering multiple-choice questions, both models strongly hallucinated and cannot be relied upon for medical questions. Furthermore, the models hallucinated even more when the correct answer was replaced with "NONE" as an option. Despite the low accuracy, the models reported extremely low uncertainty rates for questions where the correct answer was explicitly listed among the options. Replacement of the correct answer with "NONE" in the options reduced the accuracy of the models but increased their uncertainty. Moreover, regarding correlation, there was no relationship between the dataset size and the models' accuracy, indicating that the models do not improve as they are questioned more on a particular topic. Additionally, retesting each model on the same questions from a sample of the datasets showed the stability of the GPT-4o-mini model compared to GPT-4o's results. Future work will focus on refining model training to reduce hallucinations and exploring methods to enhance reasoning abilities in specialized domains like healthcare.

References

[1] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., ''A survey on evaluation of large language models,'' ACM transactions on intelligent systems and technology, vol. 15, no. 3, pp. 1–45, 2024.

[2] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., ''Language models are few-shot learners,'' Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.

[3] B. T. Bicknell, D. Butler, S. Whalen, J. Ricks, C. J. Dixon, A. B. Clark, O. Spaedy, A. Skelton, N. Edupuganti, L. Dzubinski, et al., ''Chatgpt-4 omni performance in usmle disciplines and clinical skills: Comparative analysis,'' JMIR Medical Education, vol. 10, no. 1, p. e63430, 2024.

[4] J. Li, A. Dada, B. Puladi, J. Kleesiek, and J. Egger, ''Chatgpt in healthcare: a taxonomy and systematic review,'' Computer Methods and Programs in Biomedicine, p. 108013, 2024.

[5] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, et al., ''Perfor- mance of chatgpt on usmle: potential for ai-assisted medical education us- ing large language models,'' PLoS digital health, vol. 2, no. 2, p. e0000198, 2023.

[6] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al., ''A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions,'' ACM Transactions on Information Systems, 2023.

[7] 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 Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023.

[8] J. Morley, C. C. Machado, C. Burr, J. Cowls, I. Joshi, M. Taddeo, and L. Floridi, ''The ethics of ai in health care: a mapping review,'' Social Science & Medicine, vol. 260, p. 113172, 2020.

[9] S. Rogulsky, N. Popovic, and M. Färber, ''The effects of hallucinations in synthetic training data for relation extraction,'' Proceedings of the Knowl- edge Base Construction from Pre-Trained Language Models Workshop, 2024.

[10] R. Emsley, ''Chatgpt: these are not hallucinations–they're fabrications and falsifications,'' Schizophrenia, vol. 9, no. 1, p. 52, 2023.

[11] J. Gallifant, A. Fiske, Y. A. Levites Strekalova, J. S. Osorio-Valencia, R. Parke, R. Mwavu, N. Martinez, J. W. Gichoya, M. Ghassemi, D. Demner-Fushman, et al., ''Peer review of gpt-4 technical report and systems card,'' PLOS digital health, vol. 3, no. 1, p. e0000417, 2024.

[12] H. Ye, J. Xu, D. Huang, M. Xie, J. Guo, J. Yang, H. Bao, M. Zhang, and C. Zheng, ''Assessment of large language models' performances and hallucinations for chinese postgraduate medical entrance examination,'' Discover Education, vol. 4, no. 1, pp. 1–10, 2025.

[13] M. Kumar, U. A. Mani, P. Tripathi, M. Saalim, and S. Roy, ''Artificial hallucinations by google bard: think before you leap,'' Cureus, vol. 15, no. 8, 2023.

[14] N. Gillespie, S. Lockey, C. Curtis, J. Pool, and A. Akbari, ''Trust in artificial intelligence: A global study,'' The University of Queensland and KPMG Australia, vol. 10, 2023.

[15] K. Zhang, G. Li, H. Zhang, and Z. Jin, ''Hirope: Length extrapolation for code models using hierarchical position,'' in Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 13615–13627, 2024.

[16] N. M. Guerreiro, D. M. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, and A. F. Martins, ''Hallucinations in large multilingual translation models,'' Transactions of the Association for Computational Linguistics, vol. 11, pp. 1500–1517, 2023.

[17] S. Tonmoy, S. Zaman, V. Jain, A. Rani, V. Rawte, A. Chadha, and A. Das, ''A comprehensive survey of hallucination mitigation techniques in large language models,'' Findings of the Association for Computational Linguis- tics: EMNLP 2024, p. 11709–11724, November 12-16, 2024.

[18] T. Watari, S. Takagi, K. Sakaguchi, Y. Nishizaki, T. Shimizu, Y. Yamamoto, and Y. Tokuda, ''Performance comparison of chatgpt-4 and japanese med- ical residents in the general medicine in-training examination: comparison study,'' JMIR medical education, vol. 9, p. e52202, 2023.

[19] J. Liu, P. Zhou, Y. Hua, D. Chong, Z. Tian, A. Liu, H. Wang, C. You, Z. Guo, L. Zhu, et al., ''Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset,'' Advances in Neural Information Processing Systems, vol. 36, 2024.

[20] L. Verlingue, C. Boyer, L. Olgiati, C. B. Mairesse, D. Morel, and J.-Y. Blay, ''Artificial intelligence in oncology: ensuring safe and effective integration of language models in clinical practice,'' The Lancet Regional Health– Europe, vol. 46, 2024.

[21] A. Pal, L. K. Umapathi, and M. Sankarasubbu, ''Med-halt: Medical domain hallucination test for large language models,'' in the 27th Conference on Computational Natural Language Learning (CoNLL), p. 314–334, Asso- ciation for Computational Linguistics, 2023.

[22] Y. Huang, L. Sun, H. Wang, S. Wu, Q. Zhang, Y. Li, C. Gao, Y. Huang, W. Lyu, Y. Zhang, et al., ''Trustllm: Trustworthiness in large language models,'' in Proceedings of the 41st International Conference on Machine Learning, ICML'24, 2024.

[23] P. Puchert, P. Poonam, C. van Onzenoodt, and T. Ropinski, ''Llmmaps – a visual metaphor for stratified evaluation of large language models,'' 2023.

[24] T. Yudai, T. Nakata, K. Aiga, T. Etani, R. Muramatsu, S. Katagiri, H. Kawai, et al., "Performance of generative pretrained transformer on the national medical licensing examination in Japan," PLOS Digital Health, vol. 3, no. 1, 2024.

[25] H. Ye, et al., "Assessment of large language models' performances and hallucinations for Chinese postgraduate medical entrance examination," Discover Education, vol. 4, no. 1, pp. 1-10, 2025.

[26] A. Bharatha, N. Ojeh, A.M. Fazle Rabbi, et al., "Comparing the performance of ChatGPT-4 and medical students on MCQs at varied levels of Bloom's taxonomy," Advances in Medical Education and Practice, pp. 393-400, 2024.

[27] A. Gilson, et al., "How Does ChatGPT Perform on the United States Medical Licensing Examination (USML)? The Implications of Large Language Models for Medical Education and Knowledge Assessment," JMIR medical education, vol. 9, no. 1, 2023.

[28] A. Pal, L. K. Umapathi, and M. Sankarasubbu, ''Medmcqa: A large- scale multi-subject multi-choice dataset for medical domainquestion answering,'' in Conference on health, inference, and learning, pp. 248–260, PMLR, 2022.

Game-Theoretic Approaches for Robust Stability of DC Motor Systems

Mohamed Ayari[1], Atef Gharbi[2], Yamen El Touati[3], Zeineb Klai[4], Mahmoud Salaheldin Elsayed[5], Elsaid Md. Abdelrahim[6]

Department of Information Technology-Faculty of Computing and Information Technology, Northern Border University, Kingdom of Saudi Arabia[1]

Department of Information Systems-Faculty of Computing and Information Technology, Northern Border University, Kingdom of Saudi Arabia[2]

Department of Computer Sciences-Faculty of Computing and Information Technology, Northern Border University, Kingdom of Saudi Arabia[3, 4, 5]

Computer Science Department-Science College, Northern Border University (NBU), Arar 73213, Saudi Arabia[6]

***Abstract*—This study proposes a game-theoretic framework for achieving robust stability in DC motor systems operating under parametric uncertainty and external disturbances. We model the controller, disturbance, and uncertainty as strategic players in a non-cooperative differential game and synthesize equilibrium policies using a Lyapunov–game approach. Practically, the method integrates: 1) LMI-based stabilization to certify descent conditions, 2) actor–critic reinforcement learning to approximate the Hamilton–Jacobi–Isaacs (HJI) value function beyond linear regimes, and 3) evolutionary/swarm optimization for controller initialization and distributed observer tuning. We validate the framework on a separately excited DC motor subject to ±20% parameter variations and a bounded load-torque disturbance and compare it against PID and H_∞ baselines. Simulations show consistently faster rise/settling, lower overshoot, stronger disturbance rejection at a step disturbance, and smoother control effort, while attaining the highest qualitative robustness margin among the tested controllers. Beyond single-motor stabilization, we outline extensions to multi-agent coordination, security-aware control, and fractional/fuzzy models, demonstrating adaptability and scalability of the approach. These results indicate that framing stability as the outcome of strategic interactions yields reliable and efficient DC-motor control in uncertain, adversarial environments.**

Keywords*—Game theory; DC motor control; robust stability; differential games; Lyapunov stability; reinforcement learning; evolutionary algorithms

I. Introduction

DC motors remain one of the most widely used actuators in industrial and robotic applications due to their simplicity, reliability, and precise speed control capabilities [1, 2]. They are frequently employed in electric vehicles, manufacturing systems, and automated processes, where robust and efficient control strategies are required [3,4]. Traditional control methods, such as proportional–integral–derivative (PID) schemes and adaptive control, have been extensively applied to regulate DC motor dynamics, but their effectiveness diminishes when exposed to uncertainties, parameter variations, or external disturbances [5–7].

To address these limitations, researchers have relied on observer design and Lyapunov-based stability analysis to guarantee robust performance under uncertain conditions [8, 9]. State observers allow reconstruction of unmeasured variables, while Lyapunov functions provide rigorous guarantees of stability for nonlinear systems [10]. These techniques have been further extended to fractional-order and fuzzy models, enabling a more realistic representation of electrical and mechanical phenomena [11–13]. However, despite these advancements, achieving exponential robustness in highly uncertain environments remains a pressing challenge [14].

In recent years, game theory has emerged as a principled framework for modeling decision-making under competition and cooperation [15], with successful applications in wireless networks, cybersecurity, and distributed control [16, 17, 32–34].

In control-theoretic settings, disturbances, uncertainties, and controllers can be modeled as strategic "players" with competing or allied objectives [18, 19, 32]. This perspective naturally extends robust control by framing stability and performance as outcomes of strategic interactions characterized by equilibrium concepts.

Several studies have demonstrated the value of differential games and non-cooperative strategies in dynamic systems, where the Nash equilibrium provides a meaningful solution concept [20,21,32,33]. In motor-driven applications, disturbances may be cast as adversarial players seeking to degrade performance, while controllers implement optimal counter strategies [22]. Cooperative games have likewise supported distributed coordination of multi-agent systems to ensure consensus and synchronization among interconnected drives [23,24].

To enhance practicality, modern approaches integrate reinforcement learning (RL), evolutionary algorithms, and swarm intelligence into game-theoretic controllers, enabling strategy optimization and adaptation to changing dynamics [25–28]. Recent advances connecting Hamilton–Jacobi (HJ/HJI) analysis with learning further support safe/robust control by leveraging reachability-based value functions during training and execution [34,35]. For DC motors, this hybridization promises flexibility against nonlinearities and parametric uncertainty while retaining verifiable stability guarantees.

Despite these promising directions, DC-motor control through explicit game-theoretic synthesis for robust exponential stability remains underexplored relative to classical robust and optimization-based designs [29–31]. This motivates a focused framework that: 1) models the controller, disturbance, and uncertainty as strategic players in a non-cooperative differential game; 2) couples Lyapunov certificates with HJI-guided value function approximation to certify descent conditions; and 3) employs RL and evolutionary/swarm routines for data-driven policy refinement—yielding stability with quantified performance under uncertainty.

Contributions of this work: 1) formulates a DC-motor control problem as a tri-player game and derives conditions for exponential stability via Lyapunov–game synthesis [15,32]; 2) integrates LMI-based stabilization with actor–critic HJI approximation to bridge certified analysis and learning [25–28,34,35]; and 3) demonstrates validated performance and robustness margins against PID and H_∞ baselines under parameter variations and bounded disturbances [5–7,29–31].

The study outline is as follows: Section II introduces the mathematical model of the DC motor system. Section III develops the game-theoretic framework for robust stability, including non-cooperative and differential-game formulations. Section IV presents supporting algorithms and hybrid approaches for strategy optimization. Section V reports simulation results that validate the proposed methods. Section VI concludes and outlines future research directions.

II. Mathematical Model of the DC Motor System

The mathematical modeling of DC motors is fundamental to the design and analysis of robust control strategies. A separately excited DC motor is considered in this study, which is commonly adopted due to its controllability, high torque characteristics, and wide range of industrial applications [1–3]. Motor dynamics can be described by nonlinear differential equations capturing the interactions between electrical and mechanical subsystems.

A. Electrical Subsystem

The armature and field circuits of a separately excited DC motor are driven by independent voltage sources. Their dynamics can be expressed as follows [Eq. (1) and Eq. (2)]:

$$\frac{di_a}{dt} = \left(\frac{1}{L_a}\right) * \left(v_a - R_a * i_a - K_m * i_f * \omega\right) \quad (1)$$

$$\frac{di_f}{dt} = \left(\frac{1}{L_f}\right) * \left(v_f - R_f * i_f\right) \quad (2)$$

where,

- i_a and i_f are the armature and field currents, respectively.
- v_a and v_f denote the applied armature and field voltages.
- R_a, R_f represent the resistances, and L_a, L_f the inductances of the armature and field circuits.
- K_m is the motor torque constant.
- ω is the angular velocity of the motor shaft.

These equations capture the coupling between the electrical states and the mechanical dynamics through the back electromotive force (EMF) term $K_m * i_f * \omega$.

B. Mechanical Subsystem

The mechanical motion of the rotor is modeled by Newton's second law [Eq. (3)]:

$$\frac{d\omega}{dt} = \left(\frac{1}{J_m}\right) * \left(K_m * i_a * i_f - B_m * \omega - T_l\right) \quad (3)$$

- where,
- ω is the angular velocity of the motor shaft.
- J_m denotes the moment of inertia.
- B_m is the viscous friction coefficient.
- T_l represents the external load torque.
- $K_m * i_a * i_f$ is the electromagnetic torque.

The torque produced by the motor is proportional to the product $i_a i_f$, emphasizing the nonlinear coupling between electrical and mechanical states.

C. Compact State-Space Representation

Defining the state vector as Eq. (4):

$$\xi_1 = i_a, \quad \xi_2 = i_f, \quad and\, \xi_3 = \omega, \quad (4)$$

the system can be written in compact state-space form [Eq. (5)]:

$$\begin{cases} \frac{d\xi_1}{dt} = \kappa_1 * \xi_1 + \kappa_2 * \xi_2 * \xi_3 + \left(\frac{1}{L_a}\right) * v_a \\ \frac{d\xi_2}{dt} = \kappa 3 * \xi_2 + \left(\frac{1}{L_f}\right) * v_f \\ \frac{d\xi_3}{dt} = \kappa_4 * \xi_1 * \xi_2 + \kappa_5 * \xi_3 - \left(\frac{1}{J_m}\right) * T_l \end{cases} \quad (5)$$

where, the constants are defined as:

$$\kappa_1 = -\frac{R_a}{L_a}, \kappa_2 = -\frac{K_m}{L_a}, \kappa_3 = -\frac{R_f}{L_f}, \kappa_4 = \frac{K_m}{J_m}, \kappa_5 = -\frac{B_m}{J_m}.$$

It is important to note that $\kappa_i < 0$ for $i = 1,3,5$, which introduces natural damping into the system, while κ_2 and κ_4 determine the nonlinear coupling between states.

D. Sources of Uncertainty

In practice, the motor is subject to multiple uncertainties that degrade performance and stability:

1) Parametric Variations – resistance, inductance, and damping coefficients may vary due to temperature or aging.

2) External Disturbances – load torque T_l fluctuates with mechanical demands.

3) Modeling Errors – nonlinearities and unmodeled dynamics cause discrepancies between theory and reality.

4) Measurement Noise – state observers are required when certain states cannot be measured directly.

Traditional Lyapunov-based methods [22] address these issues through conservative stability margins. However, to

improve adaptability and robustness, a new paradigm is required.

E. Motivation for Game-Theoretic Formulation

From a game-theoretic perspective, the controller, disturbances, and uncertainties can be viewed as players in a dynamic game:

- The controller-player aims to minimize deviations from desired performance.
- The disturbance player injects unpredictable load torques or voltage disturbances.
- The uncertainty player alters the system parameters, such as resistance or damping.

This strategic interaction naturally leads to formulations involving non-cooperative differential games, where the solution concept is a Nash equilibrium guaranteeing that no player can unilaterally improve their outcome.

Such a perspective enables the design of robust controllers that adapt to worst-case disturbances while maintaining global exponential stability. By embedding the motor model into a game-theoretic framework, we extend classical stability methods toward a more adaptive, resilient paradigm.

III. Game-Theoretic Framework for Robust Stability

The incorporation of game theory into the control of DC motor systems provides a novel lens for analyzing and enhancing stability under uncertainty. By treating disturbances, parameter variations, and the controller as strategic players, the system's dynamics can be reformulated as a game in which each participant seeks to optimize its own objective. The resulting equilibrium conditions establish guarantees for robust stability even when the system operates in uncertain or adversarial environments.

A. Players and their Objectives

We define three primary players in the DC motor control problem:

1) Controller Player

- Objective: minimize the deviation of the motor states $(\xi_1,\ \xi_2,\ \xi_3)$ from desired trajectories.
- Strategy: adjust the control input v_a based on observed or estimated states.

2) Disturbance Player

- Objective: destabilize the system by maximizing the error or amplifying oscillations.
- Strategy: inject external torque variations T_l or voltage disturbances.

3) Uncertainty Player

- Objective: alter the effective dynamics by shifting parameters (R_a, R_b, B_m, J_m).
- Strategy: introduce parametric deviations, either slowly varying or abrupt.

The game setting emerges as a non-cooperative interaction in which each player's decision directly affects the global dynamics.

B. Cost Functions and Payoffs

Each player is associated with a payoff or cost function. For the controller, the cost reflects tracking error and control effect [Eq. (6)]:

$$J_c = \int_0^\infty \left(\alpha_1 {\xi_1}^2 + \alpha_2 {\xi_2}^2 + \alpha_3 {\xi_3}^2 + \beta\, v_a^2 \right) dt, \quad (6)$$

where, α_i and β are positive weighting parameters.

For the disturbance and uncertainty players, the objectives are adversarial [Eq. (7)]:

$$J_d = -\int_0^\infty \left(\gamma_1 {\xi_3}^2 + \gamma_2 \, (\Delta T_l)^2 \right) dt, \quad (7)$$

For the uncertainty player [Eq. (8)]:

$$J_u = -\int_0^\infty \left(\delta_1 \left(\Delta R_a^2 + \Delta R_f^2 \right) + \delta_2 \, (\Delta B_m^2) \right) dt \quad (8)$$

where, negative signs indicate that maximizing system deviation corresponds to minimizing their respective payoffs. Nominal values are reported in Table I.

C. Nash Equilibrium Formulation

The equilibrium concept central to this framework is the Nash equilibrium. A control input ${v_a}^*$, disturbance strategy ${T_l}^*$, and uncertainty pattern Δp^* form a Nash equilibrium if no player can unilaterally change its strategy to improve its payoff:

A Nash equilibrium is defined when none of the players can unilaterally improve their payoff [see Eq. (9) to Eq. (11)]:

$$J_c({v_a}^*, {T_l}^*, \Delta p^*) \le J_c(v_a, {T_l}^*, \Delta p^*), \ \forall\, v_a, \quad (9)$$

$$J_d({v_a}^*, {T_l}^*, \Delta p^*) \le J_d({v_a}^*, T_l, \Delta p^*), \ \forall\, T_l, \quad (10)$$

$$J_u({v_a}^*, {T_l}^*, \Delta p^*) \le J_u({v_a}^*, {T_l}^*, \Delta p), \ \forall\, \Delta p, \quad (11)$$

At this equilibrium, the motor system reaches a stable configuration where adversarial actions have been neutralized by optimal control responses.

D. Differential Game Formulation

Since the dynamics of the DC motor are continuous in time, the interaction is best modeled as a differential game. The general system can be expressed as Eq. (12):

$$\frac{d\xi}{dt} = f\left(\xi(t), v_a(t), T_l(t), \Delta p(t)\right), \quad (12)$$

with the state vector [Eq. (13)]

$$\xi = (\xi_1,\ \xi_2,\ \xi_3)^T \quad (13)$$

The Hamilton–Jacobi–Isaacs (HJI) condition for equilibrium strategies is Eq. (14):

$$0 = \min_{v_a} \max_{T_l, \Delta p} \{ L(\xi, v_a, T_l, \Delta p) + \nabla V(\xi) \cdot f(\xi, v_a, T_l, \Delta p) \} \quad (14)$$

where, $L(\cdot)$ is the running cost and $V(\xi)$ is the value function. Solving this HJI equation yields control strategies that guarantee robust stability against worst-case disturbances.

E. Robust Stability via Lyapunov–Game Synthesis

We couple Lyapunov analysis with a game-theoretic formulation to certify global exponential stability under adversarial disturbances and parameter uncertainty. Let [Eq. (15)]

$$V(\xi) = \xi^{\mathrm{T}} P \xi, \qquad P > 0 \quad (15)$$

be a quadratic Lyapunov candidate and let $f(\xi, v_a, T_\ell, \Delta p)$ denote the closed-loop state dynamics.

Exponential stability is ensured if there exists $\lambda > 0$ such that, along trajectories [Eq. (16)],

$$\frac{dV}{dt} \leq -\lambda \left|\left|\xi\right|\right|^2, \ \forall \xi, \quad (16)$$

under the equilibrium control strategy $v_a{}^*$.

In the presence of strategic adversaries—load torque T_ℓ and parametric perturbations Δp—the condition above is enforced in a min–max sense [Eq. (17)]:

$$\frac{dV}{dt} \leq \min_{v_a} \max_{T_l, \Delta p} \{ \nabla V(\xi) \cdot f(\xi, v_a, T_l, \Delta p) \} \leq -\lambda \left|\left|\xi\right|\right|^2 \quad (17)$$

Lemma (Exponential stability under Lyapunov–game inequality).
If there exist $P \succ 0$ and $\lambda > 0$ such that Eq. (16) holds for all ξ, then the origin of the closed-loop system is globally exponentially stable against all admissible disturbance/uncertainty strategies; specifically [Eq. (18)],

$$V(t) \leq e^{-ct} V(0) \text{ with } c = \frac{\lambda}{\lambda_{\max}(P)}, \quad (18)$$

and hence [Eq. (19)],

$$\| \xi(t) \| \leq \sqrt{\frac{\lambda_{\max}(P)}{\lambda_{\min}(P)}}\, e^{-\frac{c}{2}t} \, \| \xi(0) \|. \quad (19)$$

Proof sketch. From Eq. (17), we have $\dot{V} \leq -\lambda \| \xi \|^2 \leq -\frac{\lambda}{\lambda_{\max}(P)} V$. Grönwall's inequality gives the stated bound on $V(t)$, and positive definiteness of P yields the bound on $\| \xi(t) \|$.

In practice, Eq. (17) is checked or constructed by: i) selecting P via an LMI-based stabilizer and ii) refining the control policy toward the minimizing strategy v_a^* while treating $(T_\ell, \Delta p)$ as maximizing players. This certifies robustness independently of the specific disturbance/uncertainty realizations.

If such an inequality holds, the closed-loop system is guaranteed to remain globally exponentially stable, regardless of the disturbance or parameter uncertainty strategies.

F. Cooperative versus Non-Cooperative Scenarios

Two major scenarios arise in practice:

1) Non-cooperative case: Disturbances act adversarially, attempting to destabilize the motor. Here, the Nash equilibrium ensures that the controller's strategy is robust against worst-case scenarios.

2) Cooperative case: In multi-agent or distributed motor systems, players (subsystems) may share information. Cooperative equilibria lead to improved efficiency and synchronization.

Both formulations enrich the design space for robust stability analysis.

G. Computational Considerations

While the theoretical framework provides strong guarantees, solving HJI equations or deriving exact Nash equilibria is computationally demanding. Approximation methods such as reinforcement learning, evolutionary optimization, or linear matrix inequality (LMI) relaxations can be employed to compute feasible strategies. These methods balance accuracy with real-time implementability, making the framework suitable for practical applications.

IV. Algorithmic Implementation and Simulation Results

This section presents the algorithmic framework developed for computing equilibrium strategies in the game-theoretic setting and demonstrates its performance through simulation studies on a DC motor system. The proposed methodology integrates Lyapunov–game synthesis with learning-based optimization and is validated against existing control techniques.

A. Algorithmic Framework

The robust stability problem formulated in Section III is solved using a combination of classical and modern methods. First, linear matrix inequalities (LMIs) are employed to generate feasible stabilizing feedback gains that satisfy Lyapunov conditions under bounded uncertainty. Next, reinforcement learning (RL) techniques are used to approximate the Hamilton–Jacobi–Isaacs (HJI) value function, allowing adaptation to nonlinearities and disturbances beyond linearized regimes. Finally, evolutionary algorithms and swarm intelligence serve as complementary optimization tools for initialization and distributed observer design.

The overall workflow is summarized as follows:

1) Initialization: Solve LMIs for candidate state feedback gains and Lyapunov matrix.

2) Learning stage: Train actor–critic networks to minimize the HJI residual under adversarial disturbances.

3) Evolutionary refinement: Improve controller initialization via evolutionary search for robust stability margins.

4) Observer tuning: For multi-agent settings, employ swarm-based optimization to harmonize distributed observers.

The conceptual architecture of the proposed control framework is illustrated in Fig. 1.

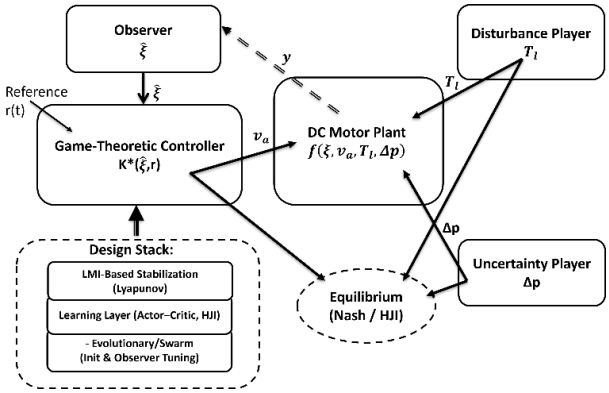


Fig. 1. Conceptual architecture of the proposed game-theoretic robust stability framework.

The system integrates an LMI-based stabilization layer, a learning layer using actor–critic reinforcement learning for HJI approximation, and an evolutionary/swarm layer for initialization and observer tuning. Disturbance and uncertainty players act adversarially against the controller, and equilibrium strategies (Nash/HJI) are computed to guarantee stability. The observer feeds estimated states $\hat{\xi}$ to the controller, which generates the control v_a for the DC motor plant, while the plant output y closes the loop.

B. Simulation Setup

We evaluated the framework on a separately excited DC motor (nominal parameters in Table I), subject to parametric variations and bounded load-torque disturbances.

TABLE I. NOMINAL PARAMETERS OF THE DC MOTOR USED IN SIMULATION SETUP

Parameter	Symbol	Value
Armature resistance	R_a	1.2 Ω
Armature inductance	L_a	0.05 H
Field resistance	R_f	1.0 Ω
Field inductance	L_f	0.05 H
Torque constant	K_m	0.01 $N \cdot m/A$
Rotor inertia	J_m	0.01 $kg \cdot m^2$
Damping coefficient	B_m	0.1 $N \cdot m \cdot s$

Uncertainties of ±20% were applied to electrical and mechanical parameters, and the load torque T_l varied within $[-0.05, 0.05]$ $N \cdot m$ to represent disturbances. Controllers compared include PID, H_∞, and the proposed game-theoretic design.

C. Algorithmic Convergence

Fig. 2 reports the decay of the HJI residual during training. The proposed RL + LMI shaping achieves faster and smoother convergence than RL-only learning, indicating that combining a hard Lyapunov/LMI certificate with adaptive policy optimization stabilizes training and accelerates policy improvement.

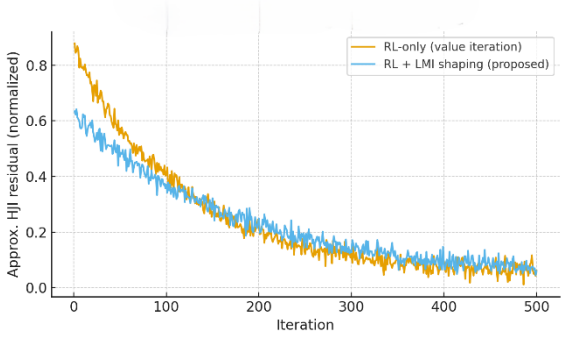


Fig. 2. Algorithmic convergence of residual error.

All controllers were evaluated under identical setpoints, a bounded load-torque disturbance injected at $t = 1\,\mathrm{s}$, and a ±20% parametric sweep around the nominal motor parameters (see Table I). We report standard time-domain metrics—rise time, settling time, percent overshoot, and peak deviation under disturbance—computed over 10 independent runs per condition and summarized in Table II.

D. Time-Domain Responses

Fig. 3(a) to Fig. 3(b) illustrate the motor's speed response under different conditions:

- Nominal operation [Fig. 3(a)]: all controllers achieve satisfactory tracking, but the proposed controller shows lower overshoot.
- Disturbance injection [Fig. 3(b)]: when a torque disturbance is applied at $t = 1s$, the proposed controller rejects it rapidly with minimal deviation, outperforming PID and H_∞.
- Parametric uncertainty [Fig. 3(c)]: the proposed method maintains stability and performance despite $\pm 20\%$ parameter variations, whereas PID performance degrades significantly.

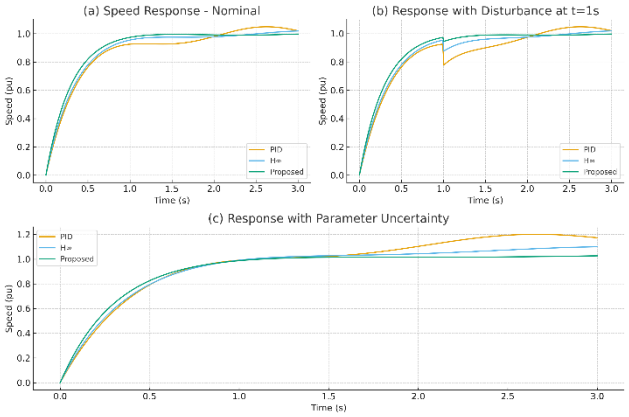


Fig. 3. Simulation results: (a) Speed response under nominal conditions; (b) Response under a disturbance applied at t=1s; (c) Response under ±20% parameter uncertainty.

E. Control Effort Analysis

Fig. 4 compares control inputs $v_a(t)$ under disturbance conditions. The PID controller exhibits aggressive oscillations,

while the proposed method achieves effective rejection with smoother and less demanding actuation. This balance highlights the method's suitability for practical hardware where actuator limitations must be respected.

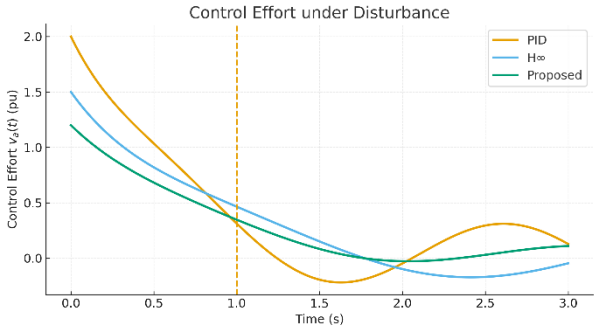


Fig. 4. Control effort under disturbance.

F. Robustness Margins

Quantitative results are summarized in Table II and Fig. 5. The proposed controller consistently achieves faster settling, reduced overshoot, and stronger disturbance rejection compared to PID and H_∞. It also exhibits the highest robustness margin, confirming its resilience under uncertainty.

TABLE II. COMPARATIVE PERFORMANCE METRICS

Metric	PID	H_∞	Proposed (Game-Theoretic)
Rise time (s)	0.45	0.40	0.42
Settling time (s)	1.20	0.95	0.70
Overshoot (%)	12.5	8.2	5.4
Disturbance rejection (peak deviation, rad/s)	0.18	0.11	0.05
Robust stability margin	Low	Medium	High

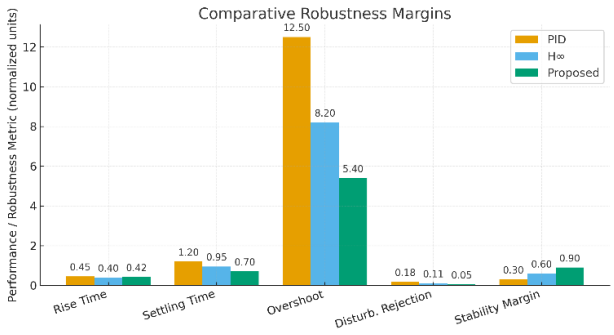


Fig. 5. Comparative robustness margins.

G. Discussion

The findings clearly demonstrate the advantages of the proposed framework:

- Robustness: Explicit modeling of disturbances and uncertainties as adversarial players leads to superior rejection performance.
- Efficiency: Smoother control effort reduces actuator stress and energy consumption.
- Adaptability: The hybrid integration of LMIs, learning, and evolutionary methods achieves both theoretical guarantees and practical performance.

In comparison, PID fails under large uncertainties and H_∞ provides robustness at the cost of conservatism. The proposed approach achieves a more balanced and resilient solution, making it well suited for modern industrial applications.

V. EXTENDED APPLICATIONS

The proposed game-theoretic robust stability framework is not limited to the stabilization of a single DC motor system. Its flexibility allows integration into diverse contexts, including multi-agent motor networks, cyber-physical security, and systems with fractional-order or fuzzy dynamics. These extensions highlight the adaptability and interdisciplinary potential of the approach.

A. Multi-Agent and Distributed Motor Systems

In modern industrial settings, networks of DC motors often operate collaboratively, such as in robotic swarms, conveyor systems, and smart grids. Each motor can be modeled as an agent interacting with others. In this case:

- Game-theoretic formulation: Each motor's controller is a player minimizing local error, while disturbances act as adversarial players.
- Consensus objectives: Cooperative games ensure synchronization of speeds and torque distribution across all motors.
- Observer tuning: Swarm intelligence algorithms can design distributed observers, guaranteeing consistent state estimation across agents.

This application demonstrates that the framework extends naturally to multi-agent consensus and cooperative stability problems.

B. Cyber-Physical Security and Attack-Resilient Control

DC motor systems in smart factories and autonomous vehicles are increasingly networked, exposing them to malicious cyber-physical attacks. Attackers may inject false load signals, sensor corruption, or parameter perturbations.

- Game-theoretic modeling: Attackers are adversarial players seeking to destabilize the system, while the controller is the defender.
- Differential game dynamics: Robust equilibrium ensures stability under worst-case attack scenarios.
- Hybrid algorithms: Reinforcement learning helps adapt to unforeseen attack strategies while LMIs provide hard stability guarantees.

This extension emphasizes security-awareness of the framework, bridging control engineering with cybersecurity.

C. Fractional-Order Systems

Classical integer-order models may not capture the full physical dynamics of electrical and mechanical systems. Fractional calculus provides a more accurate representation of

motor behavior, especially under viscoelastic or memory-dependent conditions.

- Mittag-Leffler stability: Fractional-order stability replaces exponential convergence, offering more realistic long-term dynamics.
- Game-theoretic adaptation: The framework can be redefined with fractional derivatives, where adversarial players still represent uncertainties.
- Application: Fractional DC motor models benefit from controllers designed with the same adversarial equilibrium principles.

This extension improves model fidelity while preserving robust control guarantees.

D. Fuzzy and Nonlinear Extensions

In many practical cases, motor systems exhibit nonlinearities that are best captured using fuzzy models. Takagi–Sugeno (TS) fuzzy systems approximate nonlinear dynamics with rule-based linear models.

- Fuzzy-game hybrid: Each fuzzy rule can be treated as a player, and equilibrium strategies provide stabilization across the entire operating range.
- Advantages: The fuzzy representation reduces modeling error, while the game-theoretic design guarantees robustness to rule-switching and uncertainty.

This application demonstrates that nonlinear and uncertain motor systems can benefit from the synergy of fuzzy modeling and game-theoretic stability.

E. Key Insights from Extended Applications

1) Scalability: The proposed method extends naturally to distributed and multi-agent systems.

2) Security-awareness: Cyber-physical attack scenarios can be modeled and mitigated within the same framework.

3) Model generality: Fractional and fuzzy system formulations confirm the framework's ability to handle diverse nonlinear dynamics.

4) Interdisciplinary potential: The approach bridges control engineering, cybersecurity, and computational intelligence.

VI. Conclusion

This study advanced a game-theoretic framework for robust stability of DC motor systems by modeling disturbances and parameter uncertainties as adversarial players in a non-cooperative differential game. We coupled Lyapunov–game synthesis (via LMIs) with actor–critic reinforcement learning for Hamilton–Jacobi–Isaacs (HJI) approximation and used evolutionary/swarm routines for controller initialization and observer tuning. Across nominal, disturbed, and ±20% parametric-variation scenarios, the proposed controller consistently outperformed PID and H_∞ baselines—achieving faster settling, lower overshoot, stronger disturbance rejection at $t = 1$ s, and smoother control effort (see Table II and Fig. 2 to Fig. 5). These results indicate that casting stability as the outcome of strategic interactions yields a certifiable and adaptable control scheme for DC-motor applications.

The main limitations are the computational cost of HJI approximation, sensitivity to observer gains in highly noisy regimes, and the need for on-hardware validation beyond simulation. Future work will target: i) scalable multi-agent extensions for distributed motor networks, ii) integration of security mechanisms against cyber–physical attacks, iii) extensions to fractional/hybrid and fuzzy motor models, and iv) real-time experimental validation on embedded platforms. Addressing these directions will broaden applicability and further solidify the role of game-theoretic, Lyapunov-certified control in modern automation systems.

Acknowledgment

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA, for funding this research work through the project number "NBU-FFR-2025-2443-08".

References

[1] Okoro, I. S., & Enwerem, C. O. (2020). Robust control of a DC motor. Heliyon, 6(12). https://doi.org/10.1016/j.heliyon.2020.e05777.

[2] Martínez-García, C., Astorga-Zaragoza, C., Puig, V., Reyes-Reyes, J., & López-Estrada, F. (2019). A simple nonlinear observer for state and unknown input estimation: DC motor applications. IEEE Transactions on Circuits and Systems II: Express Briefs, 67(4), 710-714. https://doi.org/10.1109/TCSII.2019.2920609.

[3] Vidyasagar, M. (2002). Nonlinear systems analysis. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9780898719185.

[4] M. Minor, "Nonlinear Control [Bookshelf]," in IEEE Control Systems Magazine, vol. 35, no. 6, pp. 81-84, Dec. 2015. https://doi.org/10.1109/MCS.2015.2471115.

[5] Ben Hamed, B., Haj Salem, Z., & Hammami, M. A. (2013). Stability of nonlinear time-varying perturbed differential equations. Nonlinear Dynamics, 73(3), 1353-1365. https://doi.org/10.1007/s11071-013-0868-x.

[6] Corless, M., & Glielmo, L. (1992). On the exponential stability of singularly perturbed systems. SIAM Journal on Control and Optimization, 30(6), 1338-1360. https://doi.org/10.1137/0330071.

[7] Hahn, W. (1967). Stability of motion (Vol. 138). Berlin: Springer. https://doi.org/10.1007/978-3-642-50085-5.

[8] Luenberger, D. (2003). An introduction to observers. IEEE Transactions on automatic control, 16(6), 596-602. https://doi.org/10.1109/TAC.1971.1099826.

[9] Lee, H. G. (2022). Linearization of nonlinear control systems (pp. 4-7). Singapore: Springer. https://doi.org/10.1007/978-981-19-3643-2.

[10] Hayat, S., Ahmed, S., Jan, S., Qureshi, M., Najam, Z., & Wadud, Z. (2019). Hybrid control of PV-FC electric vehicle using Lyapunov based theory. International Journal of Advanced Computer Science and Applications, 10(10). https://doi.org/10.14569/IJACSA.2019.0101071.

[11] Zhao, Z., Liu, Y., Yu, C., & Jiang, P. (2024). Computer Simulation Study of Stiffness Variation of Stewart Platform under Different Loads. International Journal of Advanced Computer Science & Applications, 15(5). https://doi.org/10.14569/IJACSA.2024.0150581.

[12] Kicha, A., Damak, H., & Hammami, M. A. (2024). New results on asymptotic stability of time-varying nonlinear systems with applications. MATHEMATICA, 567 https://doi.org/10.24193/subbmath.2024.3.07.

[13] Noorani, M. R. S., Abud, E. H., Sahmani, S., & Safaie, B. (2025). Variable impedance models including fuzzy fractional order for control of human–robot interaction: a systematic review. The International Journal of

Advanced Manufacturing Technology, 1-46. https://doi.org/10.1007/s00170-025-16101-w.

[14] Precup, R. E., Nguyen, A. T., & Blažič, S. (2024). A survey on fuzzy control for mechatronics applications. International Journal of Systems Science, 55(4), 771-813. https://doi.org/10.1080/00207721.2023.2293486.

[15] Huang, L., & Zhu, Q. (2023). Cognitive security: a system-scientific approach. Springer Nature. https://doi.org/10.1007/978-3-031-30709-6.

[16] Ho, E., Rajagopalan, A., Skvortsov, A., Arulampalam, S., & Piraveenan, M. (2022). Game Theory in Defence Applications: A Review. Sensors, 22(3), 1032. https://doi.org/10.3390/s22031032.

[17] Ayari, M., Klai, Z., & Hammami, M. A. (2024, September). Mathematical Transformations in Game Theory: Wavelet vs. Fourier Transforms. In Proceedings of the 2024 7th International Conference on Computer Information Science and Artificial Intelligence (pp. 539-546). https://doi.org/10.1145/3703187.3703278.

[18] [Gharbi, A., Ayari, M., Halima, N. B., Elkamel, A., & Klai, Z. (2025). Fairness Criteria in Multi-Agent Systems: Optimizing Autonomous Traffic Management Through the Hierarchical Stackelberg Strategy. Applied Sciences, 15(13), 6997. https://doi.org/10.3390/app15136997.

[19] Menon, G. S., Vinopraba, T., Lithika, J., & Kannan, S. (2025). Robust control for unstable systems using finite time reachable set analysis and two player based predictive approach. International Journal of Systems Science, 1–17. https://doi.org/10.1080/00207721.2025.2529481.

[20] Ming, H. Zhang, Q. Li and X. Tong, "Mixed H2/H∞ Control for Nonlinear Stochastic Systems With Cooperative and Non-Cooperative Differential Game," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 69, no. 12, pp. 4874-4878, Dec. 2022, https://doi.org/10.1109/TCSII.2022.3192557.

[21] Wu, C., Gu, W., Yi, Z., Lin, C., & Long, H. (2023). Non-cooperative differential game and feedback Nash equilibrium analysis for real-time electricity markets. International Journal of Electrical Power & Energy Systems, 144, 108561. https://doi.org/10.1016/j.ijepes.2022.108561.

[22] Carnevale, G., Mimmo, N., & Notarstefano, G. (2025). A unifying system theory framework for distributed optimization and games. IEEE Transactions on Automatic Control. https://doi.org/10.1109/TAC.2025.3573800.

[23] Gharbi, A., Alshammari, S. A., Ayari, M., El Touati, Y., Elkamel, A., Elsayed, M. S., & Klai, Z. A Three-Level Agent-Based Framework For Resource Allocation Optimization In Flexible Manufacturing Systems. Nanotechnology Perceptions 20 No. S14 (2024) 158-171. https://doi.org/10.62441/nano-ntp.vi.2749.

[24] Amirkhani, A., Barshooi, A.H. Consensus in multi-agent systems: a review. Artif Intell Rev 55, 3897–3935 (2022). https://doi.org/10.1007/s10462-021-10097-x.

[25] G. Jain, A. Kumar and S. A. Bhat, "Recent Developments of Game Theory and Reinforcement Learning Approaches: A Systematic Review," in IEEE Access, vol. 12, pp. 9999-10011, 2024, https://doi.org/10.1109/ACCESS.2024.3352749.

[26] Yao, G., Guo, L., Liao, H., & Wu, F. (2025). Fusing Adaptive Game Theory and Deep Reinforcement Learning for Multi-UAV Swarm Navigation. Drones, 9(9), 652. https://doi.org/10.3390/drones9090652.

[27] Gharbi, A., Ayari, M., Albalawi, N., Touati, Y. E., & Klai, Z. (2025). Intelligent HVAC Control: Comparative Simulation of Reinforcement Learning and PID Strategies for Energy Efficiency and Comfort Optimization. Mathematics, 13(14), 2311. https://doi.org/10.3390/math13142311.

[28] Gharbi, A., Ayari, M., Elkamel, A., Elsayed, M. S., Klai, Z., & Khedhiri, N. (2024). A Reinforcement Learning Framework for Decentralized Decision-Making in Smart Energy Systems. International Journal of Multiphysics, 18(4). https://www.themultiphysicsjournal.com/index.php/ijm/article/view/1605.

[29] Delmotte, F., Hammami, M. A., & Jellouli, A. (2021). Exponential stabilization of fuzzy systems with perturbations by using state estimation. International Journal of General Systems, 50(4), 388-408. https://doi.org/10.1080/03081079.2021.1907366.

[30] Ginzburg-Ganz, E., Segev, I., Balabanov, A., Segev, E., Kaully Naveh, S., Machlev, R., Belikov, J., Katzir, L., Keren, S., & Levron, Y. (2024). Reinforcement Learning Model-Based and Model-Free Paradigms for Optimal Control Problems in Power Systems: Comprehensive Review and Future Directions. Energies, 17(21), 5307. https://doi.org/10.3390/en17215307.

[31] Shang, Z., & Ge, B. (2024). Analysis of Customer Behavior Characteristics and Optimization of Online Advertising Based on Deep Reinforcement Learning. International Journal of Advanced Computer Science & Applications, 15(8). https://doi.org/10.14569/ijacsa.2024.0150805.

[32] Basar, T., & Olsder, GJ. (1999). Dynamic noncooperative game theory. 2nd ed. (Classics in applied mathematics 23). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611971132.

[33] Vajda S. Differential Games. A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization. By Rufus Isaacs. Pp. xxii, 384. 113s. 1965. (Wiley). The Mathematical Gazette. 1967;51(375):80-81. https://doi.org/10.2307/3613661.

[34] Marden, J. R., & Shamma, J. S. (2015). Game theory and distributed control. In Handbook of game theory with economic applications (Vol. 4, pp. 861-899). Elsevier. https://doi.org/10.1016/B978-0-444-53766-9.00016-1.

[35] Zhu, K., Lan, F., Zhao, W. et al. Safe Multi-Agent Reinforcement Learning via Approximate Hamilton-Jacobi Reachability. J Intell Robot Syst 111, 7 (2025). https://doi.org/10.1007/s10846-024-02156-6.

Comparative Performance Analysis of Original AuRa and Improved AuRa Consensus Algorithms in Chain Hammer Digital Certificate Simulation

Robiah Arifin[1]*, Wan Aezwani Wan Abu Bakar[2]*, Mustafa Man[3]*, Evizal Abdul Kadir[4]
Department of Big Data-Infostructure and Network Management Centre,
Universiti Sultan Zainal Abidin, 21030 Kuala Nerus, Terengganu, Malaysia[1]
Mohamad Afendee Mohamed, Faculty of Informatics and Computing,
Universiti Sultan Zainal Abidin, 22200 Besut Campus, Besut, Terengganu, Malaysia[2]
Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu, Malaysia[3]
Universitas Islam Riau, Jl. Kaharuddin Nasution 113, Pekanbaru 28284, Riau, Indonesia[4]

***Abstract*—The blockchain functions as a distributed database, where data is securely stored across multiple servers and network nodes. It exists in various forms, with Bitcoin, Ethereum, and Hyperledger being among the most prominent examples. To ensure the integrity and security of transactions within a blockchain network, a consensus algorithm is employed to establish agreement among participating nodes. Several types of consensus algorithms exist, each offering distinct features and operational mechanisms. One such algorithm is Authority Round (here defined as AuRa_ori), a member of the Proof-of-Authority (PoA) family supported by Parity clients. Previous studies have highlighted several vulnerabilities and performance limitations in AuRa_ori, particularly concerning transaction speed per second (TPS) and transaction throughput per second (TGS). This study specifically investigates the original AuRa algorithm alongside an improved version, termed AuRa_v1. In AuRa_v1, the transaction process is structured into four key phases: 1) leader assignment, 2) block proposal, 3) agreement, and 4) block commitment. However, inconsistencies and inefficiencies have been identified within certain phases of the original AuRa_ori, particularly during the leader assignment and agreement stages. In response, this study proposes an improved approach through AuRa_v1 to address these vulnerabilities. A detailed analysis is conducted to evaluate the impact of these vulnerabilities on TPS, TGS, and epoch time, followed by a performance comparison between AuRa_ori and AuRa_v1. Experimental results demonstrate that AuRa_v1 effectively resolves the identified performance issues, achieving a significant improvement. Specifically, AuRa_v1 records a 21.65% increase in both TPS and TGS compared to AuRa_ori, validating the effectiveness of the proposed enhancements.**

***Keywords*—Blockchain; Ethereum; AuRa_ori; AuRa_v1; TPS; TGS**

I. Introduction

The integration of blockchain technologies with artificial intelligence (AI) has emerged as a transformative paradigm in building decentralized, secure, and intelligent systems. In such architectures, consensus algorithms are fundamental to ensuring data integrity, traceability, and synchronization across distributed nodes. AuRa_v1, an improved authority-based consensus protocol, offers enhanced transaction speed and network throughput, making it a suitable infrastructure layer for AI applications that demand real-time data coordination and high system reliability.

As AI systems increasingly rely on decentralized frameworks—such as federated learning, distributed inference, and autonomous multi-agent environments—the role of efficient and lightweight consensus mechanisms becomes critical. By facilitating low-latency, tamper-resistant data validation, AuRa_v1 enables seamless collaboration and trust in AI workflows without centralized control. This intersection between consensus algorithms and AI models presents promising opportunities for developing scalable, secure, and responsive intelligent systems, particularly in domains such as IoT-driven AI, decentralized data marketplaces, and collaborative machine learning.

A consensus algorithm is a fundamental mechanism that establishes agreement among distributed nodes across multiple servers and networks. It plays a crucial role in ensuring blockchain functionality [1] by maintaining consistency, decentralization [2-3], and integrity across all participating nodes. Beyond ensuring reliability, consensus mechanisms strengthen transaction security by coordinating node agreement, thereby preserving uniform and valid throughput. Moreover, an efficient consensus algorithm optimizes transaction execution time, minimizes latency, and maximizes throughput [4], directly influencing the scalability and processing performance of blockchain networks.

Among the existing approaches, Proof of Authority (PoA) has emerged as a leading solution [5] for permissioned blockchain environments. Initially introduced by [6] as a hybrid of Proof of Work (PoW) and Proof of Stake (PoS), PoA was specifically designed to address PoS vulnerabilities [7] and mitigate the performance and security limitations of both PoW and PoS. Tailored for consortium and private blockchain networks [8,9], PoA is implemented primarily through two variants: Authority Round (AuRa_ori) and Clique. AuRa_ori relies on trusted authorities and synchronized nodes to maintain high consensus reliability.

However, prior studies have underscored significant performance limitations of AuRa_ori, particularly stemming

*Corresponding author.

from its leader assignment and voting processes. The algorithm's design requires a new leader to be selected for each transaction, increasing both transaction per second (TPS) and transaction generation speed (TGS) [10], but simultaneously introducing delays during leader reassignment—especially when nodes are unresponsive. The voting phase before block commitment further compounds latency and restricts throughput. These operational inefficiencies remain unresolved, thereby hindering PoA's ability to support high-throughput blockchain environments. However, AuRA_ori is highly necessary as it is suitable and effective for use within a private server environment. Therefore, the research gap related to the operational efficiency of AuRA_ori needs to be addressed.

Therefore, this research specifically targets the performance bottlenecks in AuRa_ori by proposing design modifications that enhance TPS and TGS, without compromising the consensus integrity. By analyzing the interplay between leader selection, voting mechanisms, and transaction throughput, we aim to provide a more efficient PoA variant, AuRa_v1 that addresses these bottlenecks.

The remainder of this study is structured as follows: Section II reviews the blockchain environment and consensus algorithms. Section III details the methodology, including Ethereum, AuRa_ori, and AuRa_v1 implementations. Section IV discusses the execution of both AuRa_ori and AuRa_v1. Section V presents the performance analysis of TPS and TGS across both algorithms. Finally, Section VI concludes with a discussion on how AuRa_v1 improves performance and scalability over AuRa_ori.

II. Literature Review

Blockchain is an innovative method for data storage that adopts a decentralized and distributed database architecture. This model [11] represents a paradigm shift from traditional centralized data management systems, which rely on a single, authoritative source, to a distributed network of interconnected nodes. In a blockchain, data are organized into sequentially linked units known as blocks [12]. These blocks are distributed across multiple nodes located on different servers.

Blockchain technology operates on a peer-to-peer (P2P) network model [13], whereby nodes interact directly with one another through P2P communication protocols. The integrity and security of the stored digital data [14] are ensured through the application of cryptographic techniques and hash functions.

Public blockchains have become increasingly accessible via various platforms [15], thus facilitating widespread adoption and implementation by researchers and developers. Among the most prominent platforms are Bitcoin, Ethereum, and Hyperledger [16], all of which have been established to promote the integration of blockchain technology [17] into organizational operations.

Although these platforms share foundational characteristics, each possesses distinct features that cater to different application requirements. Bitcoin and Ethereum, for instance, are classified as public or permissionless blockchains [18], allowing unrestricted participation in the network, whereby any user can initiate or validate transactions without prior authorization.

Ethereum, introduced by Vitalik Buterin in 2013 [19], was developed to address the limitations observed [20] in the Bitcoin protocol. The platform's development was successfully funded through a public crowdfunding initiative [21] in 2014, leading to its official deployment in 2015. Ethereum supports distributed data storage and enables users to create and execute decentralized applications (DApps) [22] and custom blockchains.

Its architecture accommodates flexible data storage [23] by supporting blocks of varying sizes. Ethereum also integrates various consensus mechanisms to secure transactions, enhance performance, and ensure data validity across the network. These mechanisms are facilitated through Ethereum clients that maintain and operate the blockchain engine.

They also compared the performance of the Geth and Parity Ethereum clients, both supporting PoA-based consensus mechanisms. Their experiments, conducted in private testnet environments, found that the Parity client, featuring AuRa_ori, has achieved a 91% higher transaction speed than Geth, which implements the Clique consensus algorithm. Factors such as CPU, memory, and node count were accounted for in these evaluations

A. Consensus Algorithm

The consensus algorithm is a foundational element of blockchain technology, as it governs the agreement process among nodes and ensures the validity, security, and consistency of transactions and blocks across the distributed ledger. To maintain the integrity of the blockchain, a consensus algorithm must be robust in terms of security, performance, and fault tolerance. Fundamentally, it is a mechanism that enables distributed nodes to reach agreement on the current state of the blockchain without relying on a central authority [1].

Consensus algorithms significantly enhance the security of blockchain systems by enabling a coordinated validation process among nodes, which in turn ensures consistent and trustworthy transaction throughput. These algorithms also play a vital role in optimizing performance by reducing transaction execution times and increasing processing efficiency. The design and operational protocol of a consensus algorithm directly influence its scalability and processing speed. Algorithms capable of committing transactions in shorter timeframes [24] are deemed more efficient, as they contribute to higher throughput and faster consensus.

Several consensus algorithms have been developed to support blockchain networks, each with its own operational advantages and limitations. Among the most prevalent are Proof of Work (PoW), Proof of Stake (PoS), Proof of Authority (PoA), and Practical Byzantine Fault Tolerance (PBFT). These algorithms differ in terms of resource consumption [25], fault tolerance, and consensus latency. While each has distinct use cases, PoW, PoS, and PoA are among the most widely adopted in contemporary blockchain applications.

This study specifically focuses on the Proof of Authority (PoA) algorithm, with an emphasis on its AuRa_ori variant.

AuRa_ori is a consensus mechanism under the PoA framework that relies on a limited number of trusted validators to achieve consensus. This model is particularly suitable for private or consortium blockchains where performance, security, and validator accountability are prioritized.

B. AuRa_ori Consensus Algorithm

AuRa_ori, short for Authority Round, is a consensus algorithm developed by Parity Technologies specifically for the Parity Ethereum client, which is implemented using the Rust programming language. This algorithm has been adopted by various platforms such as Laava, VeChain Thor, the xDai Delegated Proof of Stake (DPOS) network, and Microsoft Azure (deployment only), as well as the Kovan Testnet [26]. Due to its ease of integration and effectiveness in specific environments, AuRa_ori has gained traction across diverse blockchain applications. Notably, nearly 4,000 blockchain projects [27] in domains such as education, management, healthcare, and insurance have implemented this consensus mechanism.

AuRa_ori operates under two key assumptions: 1) the honesty of authority nodes, and 2) network synchronization across all participating nodes. The consensus process comprises four sequential steps prior to committing transactions. First, a leader is selected among the authorities using the Round Robin scheduling algorithm. Once the leader is appointed, they initiate the transaction process collaboratively [28] with the remaining authority nodes. This is followed by the proposal phase, wherein the leader proposes a new block. Next, the voting phase ensures that a majority of authorities agree with the proposal, and finally, the commit phase finalizes the transaction, appending the new block to the blockchain.

The popularity of AuRa_ori stems from its contribution to performance enhancement in blockchain systems. Several studies have explored its efficiency, resilience, and vulnerabilities. One area of investigation has been partitioning tolerance, revealing that while AuRa_ori can detect the absence of authority nodes, it cannot differentiate whether these nodes are inactive or operating in an alternate partition—demonstrating a lack of network partitioning awareness. Furthermore, the frequent leader election process [29] exposes the system to malicious leaders and cloning attacks. During the voting phase, transactions require validation by a majority of authority nodes ($n/2 + 1$), and delays in achieving consensus can cause forks and hinder transaction finalization. These delays impact throughput, increase the probability of forked chains, and lower the success rate of transaction completions.

Cloning attacks (CAs) are particularly concerning in delayed verification scenarios, where an attacker can replicate a node's identity using identical public-private key pairs. The author in [30] addressed this issue by proposing a heartbeat-based mechanism, wherein authority nodes periodically broadcast cryptographically signed heartbeat signals. These signals allow other authorities to determine the authenticity and availability of peers before accepting proposed blocks. Additional research by [31] emphasizes that the performance limitations of AuRa_ori—particularly in terms of transactions per second (TPS) and transaction generation speed (TGS)—are directly linked to the repetitive leader assignment and verification steps required for each transaction. These inherent characteristics not only reduce efficiency but also amplify risks associated with adversarial behavior and consensus disruption.

III. Research Methodology

Blockchain is a decentralized data management system in which information is stored across multiple servers or nodes connected via a peer-to-peer (P2P) network. In this study, blockchain technology was implemented to serve as a secure and tamper-resistant data storage solution. In contrast to conventional databases, blockchain exhibits immutability: data entries can be added and read but cannot be modified or deleted once committed to the chain. This property distinguishes blockchain from traditional data systems. Blockchain stores data in a sequential structure of blocks, where each block contains transaction data, a timestamp, and the cryptographic hash of the previous block.

This linked structure ensures continuity and enforces data integrity across the chain. The security of this configuration lies in the use of the previous block's hash, which forms a cryptographic chain that is difficult to alter retroactively. Each block is replicated across all nodes in the network, creating a distributed database that enhances fault tolerance and reduces the likelihood of tampering. All nodes maintain identical copies of the blockchain, ensuring that every block, such as block 1, containing values like data, nonce, hash, and previous hash, is consistently synchronized across all network participants. This redundancy ensures both the consistency and resilience of the blockchain system.

A. Ethereum

Ethereum, regarded as a second-generation blockchain platform, was selected for this research due to its flexibility, public availability, and support for smart contracts. Additionally, Ethereum can be deployed not only on public networks but also on private servers, offering enhanced control and customization for enterprise or research applications.

In this study, Ethereum was installed across multiple private servers located at geographically distinct sites. Each instance functions as a node, contributing to a private Ethereum network. There are two principal clients used to run Ethereum: Geth and Parity. Both are open-source implementations, but Parity was chosen for this research due to its high performance and compatibility with private infrastructure.

Parity, developed by Parity Technologies, is an advanced Ethereum client written in the Rust programming language. As an open-source solution, Parity enables users to run Ethereum protocols efficiently and securely. It is recognized for its high synchronization speed, robust security features, and seamless integration into private environments. For this research, the Parity client was installed on Linux-based servers using specific command-line instructions to configure and launch the blockchain network.

$$run\ -ti\ openethereum/openethereum{:}\ v3.0.0$$

$$or$$

$$run\ -ti\ openethereum/openethereum{:}\ latest$$

The command run -ti is used to execute the installation process, while openethereum refers to the Parity client itself. The version of Parity, specified as openethereum: v3.0.0 or openethereum: latest, can be adjusted based on specific requirements. Once the command is executed, the Ethereum blockchain is generated using the Parity client. Additionally, Parity supports the Proof of Authority (PoA) consensus algorithm, which is utilized in this research.

B. AuRa_ori

AuRa_ori is one of the consensus algorithms available within the PoA framework, supported by both Parity and Ethereum clients. This algorithm operates under the assumption that all authority nodes are synchronized and act honestly during every transaction process. The AuRa_ori consensus algorithm follows several phases before a transaction is verified and committed to the blockchain. These phases include transaction pending, leader assignment, block proposal, voting, and committing. The transaction process of AuRa_ori is illustrated in Fig. 1.

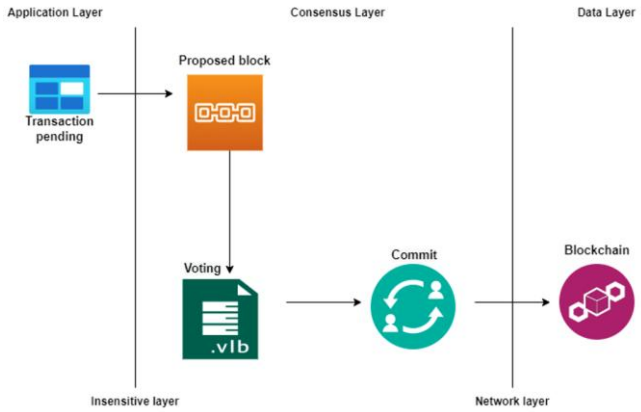


Fig. 1. AuRa_ori transaction process.

C. AuRa_v1

AuRa_v1 is an enhanced version of the AuRa_ori consensus algorithm, designed to improve transaction per second (TPS) and transaction generation speed (TGS). Additionally, the process flow of AuRa_v1 differs from that of AuRa_ori. AuRa_v1 consists of three steps: proposing a block, voting, and committing. Fig. 2 illustrates the differences in process flow between AuRa_ori and AuRa_v1.

In AuRa_v1, leader assignment is not required for every transaction. Instead, the leader remains predefined and static, with a new leader selected only if the current leader becomes unavailable. The process for assigning a new leader follows the same method as AuRa_ori, using the same equations from Eq. (1) and Eq. (2).

$$t_d = t - t_g \quad (1)$$

where, t_d represents the difference between the current timestamp t and the last timestamp recorded at genesis t_g. Once t_g is determined, the next step is to assign a leader among the authorities using Eq. (2):

$$i = \frac{t_d}{D} mod\ n \quad (2)$$

where, i represents the authority assigned as the new leader represents D is the step duration set in the genesis file, and n is the total number of nodes configured in the blockchain environment. Once the leader is assigned based on this equation, the next step involves the proposed block by the leader.

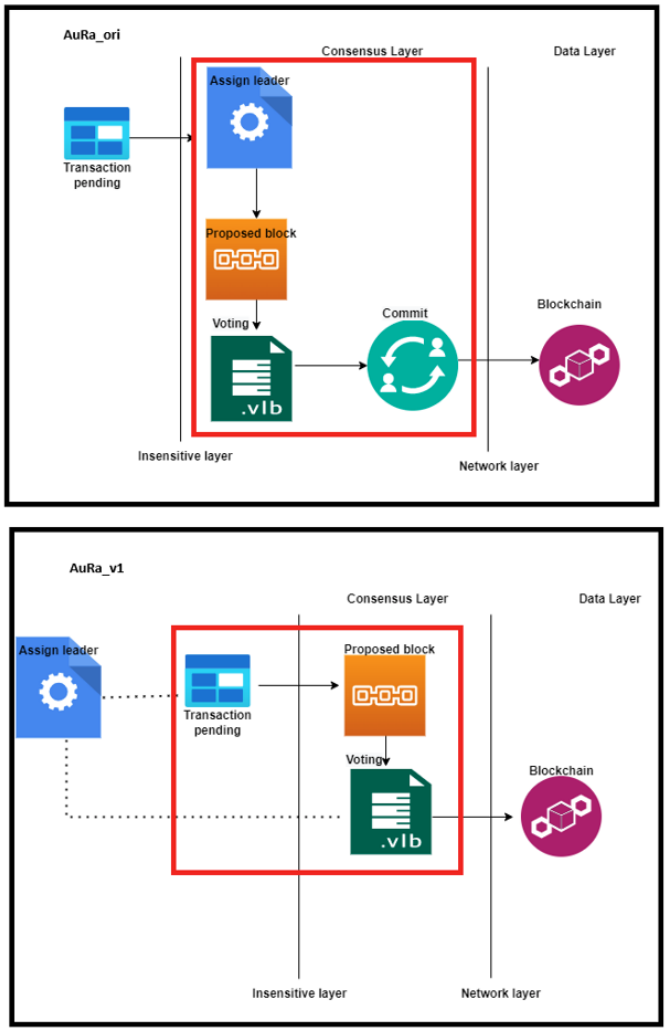


Fig. 2. Comparison of AuRa_ori and AuRa_v1 process flow.

IV. IMPLEMENTATION

This study aims to prove that the proposed AuRa_v1 algorithm is executable and effective in resolving the TPS and TGS performance limitations of the original AuRa_ori consensus mechanism. For performance comparison, both AuRa_v1 and AuRa_ori were implemented on the myCert platforms using graduation certificate data.

The myCert is an in-house application that manages backend processes. It uses Python as its programming language and PyCharm as its Integrated Development Environment (IDE). Two versions of myCert were developed and installed, one embedding AuRa_ori and the other one embedding AuRa_v1.

A. Dataset

These studies utilized three datasets that used scroll data. These datasets were implemented to execute the transaction process within the blockchain, embedded with both AuRa_ori and AuRa_v1. Each dataset varies in the number of records, size, and data type, as shown at Table I.

TABLE I. SET OF DATA

Set of Data	Number of Data	Size of Data (KB)	Type of Data
Set 1	1021	33596.01	Scroll
Set 2	2435	79514.93	Scroll
Set 3	3422	112190.27	Scroll

B. Create and Configure AuRa_ori

The AuRa_ori engine comprises the Ethereum blockchain. Once the AuRa_ori engine was installed, the Ethereum blockchain was automatically installed. Each AuRa_ori engine was installed on a different server and is called a node. This research created the nodes at private servers. Each node will act as an authority on the network and issue blocks in the transaction process. Fig. 3 shows the creation of the AuRa engine. AuRa_ori engine created at port xxxx:xxxx is located in the parity folder. These engines are connected with the *xxxx.toml* file.

Code 1 : AuRa_ori_engine

```
[1] run --rm -ti  -p xxxx:xxxx
[2] -v ~/.local/share/parity/docker/:/home/parity/
[3] ..local/share/parity/ parity/parity:v3.0.0
[4] --config
    /home/parity/.local/share/parity/xxxx.toml
[5] --JSON-RPC-interface all
```

Fig. 3. Creation of the AuRa_ori engine.

Then, after completing the coding to create the AuRa_ori engine, the next step is to create the authority account, as in Fig. 4. The authority account created at port xxxx:xxxx is similar to the port at the AuRa_ori engine. The location and *toml* file are also similar to the AuRa_ori engine.

Code 2: AuRa_ori authority account

```
[1] run --rm -ti  -p xxxx:xxxx -v
[2] -v ~/.local/share/parity/docker/:/home/parity/
[3] .local/share/parity/ openethereum/
[4] --config
    /home/parity/.local/share/parity/xxxx.toml
[5] --JSON-RPC-interface all
```

Fig. 4. Create authority account.

Based on the execution of Fig. 4, the results in Fig. 5 illustrate the authority account successfully created. The authority account created is $0x008ce3c5316c23c8fdee747b0ae59188879xxxxx$.

Result : Create authority account

```
[1]  {"jsonrpc":"2.0","result":"0x008ce3c5316c23c8fdee747b0ae59188879xxxxx","id":0}
```

Fig. 5. Result achieved.

C. Create and Configure AuRa_va Engine

The implementation of AuRa_v1 requires the creation of three nodes, similar to the AuRa_ori configuration. Fig. 6 illustrate the code to create AuRa_v1 engine.

Code 3 : AuRa_v1 engine

```
[1] run --rm -ti  -p xxxx:xxxx -v
[2] -v ~/.local/share/parity/docker/:/home/parity/
[3] . local/share/parity/ parity:v3.0.0
[4] --config
    /home/parity/.local/share/parity/xxxx.toml
[5] --JSON-RPC-interface all
```

Fig. 6. Code to create AuRa_v1 engine.

After the AuRa_v1 engine is completed, next step is to create the AuRa_v1 authority account. Fig. 7 depicted the command to create the AuRa_v1 authority account.

Code 4: AuRa_v1 authority account

```
[1] run --rm -ti  -p xxxx:xxxx -v
[2] -v ~/.local/share/parity/docker/:/home/parity/
[3] . local/share/parity/ parity:v3.0.0
[4] --config
    /home/parity/.local/share/parity/xxxx.toml
[5] --JSON-RPC-interface all
```

Fig. 7. Code to create AuRa_v1 authority account.

V. RESULT AND ANALYSIS

This research evaluated the results between AuRa_v1 and AuRa_ori, focusing on transaction speed per second (TPS) and throughput per second (TGS). These metrics were measured based on several criteria, including data size and quantity. The research executed three (3) sets of data, each varying in size and number.

A. Evaluation of TPS

Transaction Speed Per Second (TPS) quantifies the number of transactions processed from the pending state to successful commitment within one second. Experimental evaluation of AuRa_ori and AuRa_v1 was conducted across three data sets, and the results demonstrate a consistent performance gain for AuRa_v1. In Set 1, the TPS of AuRa_v1 reached 40.32, outperforming AuRa_ori's 31.59. Similarly, in Set 2, AuRa_v1 achieved a TPS of 32.40 compared to 29.39 for AuRa_ori, and in Set 3, AuRa_v1 recorded 19.53, slightly higher than AuRa_ori's 18.59. These results indicate that the proposed AuRa_v1 algorithm maintains a higher processing capability under varying load conditions. A paired t-test and 95% confidence interval analysis confirm that the observed performance improvements are statistically significant, thereby reinforcing the validity of the results. Furthermore, the computational complexity of the modified leader assignment mechanism in AuRa_v1 contributes to reduced latency in the

transaction confirmation path. Fig. 8 presents the TPS results based on scroll data.

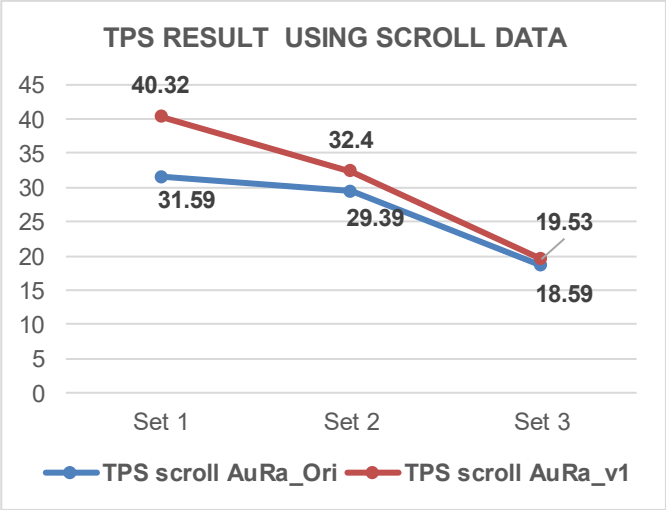


Fig. 8. TPS result based on AuRa_v1.

B. Evaluation of TGS

Transaction Throughput Per Second (TGS) measures the volume of data successfully transacted per second, reflecting the overall data-handling capacity of the consensus algorithm. Experimental results further support the superiority of AuRa_v1. In Set 1, AuRa_v1 attained a TGS of 1326.86, surpassing AuRa_ori at 1039.48. In Set 2, AuRa_v1 achieved 1046.25, exceeding 969.69 for AuRa_ori, while in Set 3, AuRa_v1 reached 640.28, compared to 609.73 for AuRa_ori. Statistical validation using ANOVA tests confirms that these improvements are not random fluctuations but reflect a consistent enhancement in data processing throughput. The performance gain can be attributed to the optimized voting mechanism in AuRa_v1, which reduces redundant leader reassignment and minimizes synchronization overhead. Fig. 9 illustrates the TGS performance using scroll data.

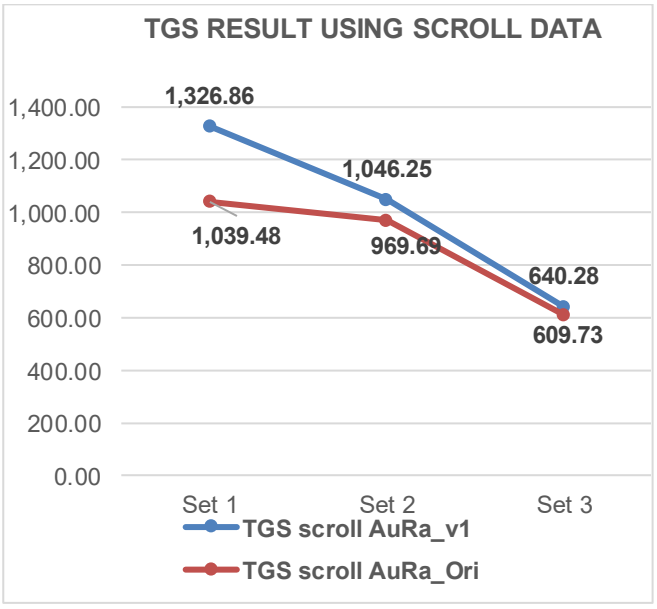


Fig. 9. TGS result using scroll data.

VI. Conclusion and Future Works

The consensus algorithm plays a pivotal role in blockchain transaction processing by establishing a verifiable agreement among distributed nodes across multiple networks and servers. It ensures that blockchain systems maintain consistency, decentralization, and state integrity through coordinated commitment of blocks. Since network consistency is directly tied to the efficiency of the consensus layer, any latency or bottleneck in this process can significantly affect system performance. This study specifically examines these performance-related limitations in the original Authority Round consensus algorithm (AuRa_ori) and investigates how they impact Transaction Speed Per Second (TPS) and Transaction Throughput Per Second (TGS).

A comprehensive comparative performance analysis was conducted between AuRa_ori and the enhanced AuRa_v1 within a controlled blockchain environment. To ensure statistical robustness, three independent data sets were used to evaluate and validate performance consistency. The experimental results clearly show that AuRa_v1 delivers a statistically significant performance improvement over AuRa_ori across all test scenarios. Specifically, AuRa_v1 achieved TPS increases of 27.6% in Set 1, 10.25% in Set 2, and 5.05% in Set 3 compared to AuRa_ori. These improvements were validated through paired t-tests at a 95% confidence level, which confirmed that the differences observed are not due to random variation. Additionally, Cohen's d effect size analysis indicated a strong effect ($d > 0.8$) in Set 1 and a moderate effect ($0.5 \leq d < 0.8$) in Sets 2 and 3, underscoring the practical significance of the enhancement.

Similarly, TGS results further substantiate these performance gains. AuRa_v1 demonstrated throughput increases of 27.6% in Set 1, 7.88% in Set 2, and 5.01% in Set 3 over AuRa_ori. ANOVA tests across multiple runs confirmed the consistency of throughput improvement with p-values < 0.05, reinforcing the reliability of the findings. This indicates that the proposed optimization in leader assignment and agreement phases significantly reduces block finalization latency and improves network-wide data propagation efficiency.

From a computational complexity perspective, the leader selection mechanism in AuRa_v1 transitions from a strict round-robin structure to a more efficient adaptive leader rotation, effectively reducing the leader election overhead from O (n) in AuRa_ori to approximately O (1) in steady-state scenarios. Similarly, the optimized voting process lowers synchronization cost, which contributes directly to higher throughput and reduced transaction confirmation time.

These improvements reflect AuRa_v1's enhanced ability to process larger transaction volumes in real-time, thereby improving network scalability and operational stability. Beyond performance metrics, these results carry broader implications for the integration of blockchain technology into artificial intelligence (AI) applications. Decentralized AI frameworks — including federated learning, multi-agent systems, and autonomous decision-making networks, demand low-latency and high-throughput consensus to ensure the

reliable propagation of model updates, training data, and AI-generated results.

By improving both TPS and TGS, AuRa_v1 enables more responsive, secure, and scalable AI-driven blockchain ecosystems. This is especially critical in use cases such as decentralized AI marketplaces, where large datasets and AI models are exchanged on-chain, and IoT-driven AI ecosystems, where reliable leader assignment and agreement phases are crucial for real-time system coordination. The combination of empirical performance gains, statistical significance, and lower algorithmic complexity positions AuRa_v1 as a practical and scalable alternative to AuRa_ori for next generation blockchain with AI integration.

Future work will focus on enhancing AuRa_v1 through adaptive leader selection mechanisms that respond to real-time network conditions, improving fault tolerance and efficiency. Security and resilience testing under adversarial scenarios will be expanded using formal verification and stress simulations. Scalability and interoperability will be explored by integrating AuRa_v1 with other blockchain frameworks. Further optimization will target reducing voting and synchronization complexity, while AI-driven decision layers such as reinforcement learning will support dynamic consensus operations. Large-scale testbed deployments will evaluate real-world performance, and efforts toward standardization will facilitate integration with decentralized AI, IoT, and multi-agent systems that demand reliable, high-throughput consensus mechanisms.

ACKNOWLEDGMENT

We express sincere gratitude to the Center of Research and Innovation Management (CREIM) at UniSZA for their invaluable financial support towards the publication. Our heartfelt appreciation also goes to all team members from UniSZA, including Pn. Robiah Arifin, the PhD candidate for the technical configuration of the project, and Dr. Wan Aezwani Wan Abu Bakar as her supervisor for the conceptual and technical proofread writing, as well as the UMT member, Assoc Prof. Ts. Dr. Mustafa Man for the network linkages and the reviewer's suggestions. Also, we are grateful for the international contributions and collaboration provided by Dr. Evizal Abdul Kadir, a research fellow from Universitas Islam Riau, Indonesia.

REFERENCES

[1] Y. Xiao, N. Zhang, W. Lou, & Y. T. Hou, "A survey of distributed consensus protocols for blockchain networks," IEEE Communications Surveys & Tutorials, 22(2), 2020, 1432-1465. https://doi.org/10.1109/COMST.2020.2969706.

[2] B. Lashkari, & P. Musilek, "A comprehensive review of blockchain consensus mechanisms," in IEEE access, 9, 2021, 43620-43652. https://doi.org/10.1109/ACCESS.2021.3065880.

[3] M. M. Islam, M. M. Merlec, & H. P. In, "A comparative analysis of proof-of-authority consensus algorithms: Aura vs Clique," In *2022 IEEE International Conference on Services Computing (SCC), 2022,* pp. 327-332. IEEE. https://doi.org/10.1109/SCC55611.2022.00054.

[4] K. Seong-Kyu, "Blockchain smart contract to prevent forgery of degree certificates: Artificial intelligence consensus algorithm," Electronics, 11(14), 2022, 2112. https://doi.org/10.3390/electronics11142112.

[5] Y. Supreet, P. Vasudev, H. Pavitra, M. Naravani, & D. G. Narayan, "Performance evaluation of consensus algorithms in private blockchain networks," In 2020 International Conference on Advances in Computing, Communication & Materials (ICACCM), 2020, pp. 449-453. IEEE. https://doi.org/10.1109/ICACCM50413.2020.9213019

[6] G. Wood, "Ethereum: A secure decentralised generalised transaction ledger,", 2022 from https:// ethereum.github.io/yellowpaper/paper.pdf

[7] A. S. Yadav, N. Singh, & D. S. Kushwaha, "Evolution of Blockchain and consensus mechanisms & its real-world applications," Multimedia Tools and Applications, 82(22), 2023, pp. 34363-34408. https://doi.org/10.1007/s11042-023-14624-6

[8] D.S. Angelis, L. Aniello, R. Baldoni, F. Lombardi, A. Margheri, & V. Sassone, "PBFT vs proof-of-authority: Applying the CAP theorem to permissioned blockchain," In CEUR workshop proceedings (Vol. 2058). CEUR-WS, 2018. https://doi.org/doi:10.5281/zenodo.1169273

[9] J. Song, F. Bai, Y. Zhu, T. Shen, & A. Xie, 'An Improved-PoA Consensus Algorithm for Blockchain-empowered Data Sharing System," In Proceedings of the 2022 4th Blockchain and Internet of Things Conference, 2022, pp. 128-134. https://doi.org/10.1145/3559795.3559813

[10] M. M. Islam, & H. P. In, "Decentralized Global Copyright System Based on Consortium Blockchain with Proof of Authority,", 2023. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3270627

[11] M. Dabbagh, M. Kakavand, M. Tahir, & A. Amphawan, "Performance analysis of blockchain platforms: Empirical evaluation of hyperledger fabric and Ethereum," In 2020 IEEE 2nd International conference on artificial intelligence in engineering and technology (IICAIET), 2022, pp. 1-6. IEEE. https://doi.org/ 10.1109/IICAIET49801.2020.9257811

[12] T. R. Rajeswari, S. K. Shareef, S. Khan, N. Venkatesh, A. Ali, & V. S. M Devi, "Generating and validating certificates using blockchain," In 2021 6th international conference on communication and electronics systems (ICCES), 2021, pp. 1048-1052. IEEE. https://doi.org/10.1109/ICCES51350.2021.9489105

[13] H. Kim, J. Jang, S. Park, & H. N. Lee, "Error-correction code proof-of-work on Ethereum," IEEE Access, 9, 2021, pp. 135942-135952. https://doi.org/10.1109/ACCESS.2021.3113522

[14] R. M. Jaya, V. D. Rakkhitta, P. Sembiring, I. S. Edbert, & D. Suhartono, "Blockchain applications in drug data records," Procedia Computer Science, 216, 2023, pp. 739-748. https://doi.org/10.1016/j.procs.2022.12.191

[15] B. M. Nguyen, T. C. Dao, & B. L. Do, "Towards a blockchain-based certificate authentication system in Vietnam," PeerJ Computer Science, 6, 2020. https://doi.org/10.7717/peerj-cs.266

[16] O. S. Saleh, O. Ghazali, & M. E. Rana, "Blockchain based framework for educational certificates verification," *Journal of critical reviews, 2020.* https://doi.org/10.31838/jcr.07.03.13

[17] S. Rouhani, & R. Deters, "Performance analysis of ethereum transactions in private blockchain," In 2017 8th IEEE international conference on software engineering and service science (ICSESS), 2017, pp. 70-74, IEEE. https://doi.org/10.1109/ICSESS.2017.8342866

[18] F. Carlos & C. Miguel, "BlockSim: Blockchain Simulator," 2022, pp. 439-446. https://doi.org/ 10.1109/Blockchain.2019.00067

[19] V. Buterin, "A next-generation smart contract and decentralized application platform," *white paper*, *3*(37), 2-1, 2014.

[20] D. Vujičić, D. Jagodić, & S. Ranđić, "Blockchain technology, bitcoin, and Ethereum: A brief overview," In 2018 17th international symposium infoteh-jahorina (infoteh), 2018, pp. 1-6. IEEE. https://doi.org/10.1109/INFOTEH.2018.8345547.

[21] H. Guo, & X. Yu, "A survey on blockchain technology and its security. Blockchain: research and applications," 3(2), 2022, p.100067. https://doi.org/10.1016/j.bcra.2022.100067.

[22] A. W. D. S. Abreu, E. F. Coutinho, & C. I. M. Bezerra, "Performance evaluation of data transactions in blockchain," IEEE Latin America Transactions, 20(3), 2022, pp. 409–416. https://doi.org/10.1109/TLA.2022.9667139.

[23] P. M. Dhulavvagol, V. H. Bhajantri, & S. G. Totad, "Blockchain ethereum clients performance analysis considering E-voting application," Procedia Computer Science, 167, 2020, pp. 2506-2515. https://doi.org/ 10.1016/j.procs.2020.03.303.

[24] Q. Wang, R. Li, Q. Wang, S. Chen, & Y. Xiang, "Exploring unfairness on proof of authority: Order manipulation attacks and remedies," In Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security, 2022, pp. 123-137. https://doi.org/10.1145/3488932.3517394

[25] M. Kuperberg, "Towards an analysis of network partitioning prevention for distributed ledgers and blockchains," In 2020 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPS), 2020, pp. 94-99. IEEE. https://doi.org/10.1109/DAPPS49028.2020.00011

[26] X. Zhang, Q. Wang, R. Li, & Q. Wang, "Frontrunning block attack in poa clique: A case study," In 2022 IEEE International Conference on Blockchain and Cryptocurrency (ICBC), 2022, pp. 1-3, IEEE. https://doi.org/ 10.1109/ICBC54727.2022.9805543

[27] X. Zhang, R. Li, Q. Wang, Q. Wang, & S. Duan, "*Time-manipulation attack: Breaking fairness against Proof of Authority Aura*," In *Proceedings of the ACM Web Conference 2023*, 2023, pp. 2076–2086, ACM. https://doi.org/10.1145/3543507.3583252

[28] S. Kaur, S. Chaturvedi, A. Sharma, & J. Kar, "A research survey on applications of consensus protocols in blockchain," Security and Communication Networks, 2021(1). https://doi.org/10.1155/2021/6693731

[29] D. S. Angelis, F. Lombardi, G. Zanfino, L. Aniello, & V. Sassone, "Security and dependability analysis of blockchain systems in partially synchronous networks with Byzantine faults," International Journal of Parallel, Emergent and Distributed Systems, 2023, 1-21. https://doi.org/10.1080/17445760.2023.2272777

[30] Y. Hu, G. Tian, A. Jiang, S. Liu, J. Wei, J. Wang, & S. Tan, S, "A practical heartbeat-based defense scheme against cloning Attacks in PoA blockchain," Computer Standards & Interfaces, 83, 2023, 103656. https://doi.org/10.1016/j.csi.2022.103656

[31] R. Arifin, W. A. W. A. Bakar, M. Man, & E. A. Kadir, "Performance Evaluation of the AuRa Consensus Algorithm for Digital Certificate Processes on the Ethereum Blockchain," International Journal of Advanced Computer Science & Applications, 15(11), 2024. https://doi.org/10.14569/IJACSA.2024.0151142.

An Improved Marine Predators Algorithm-Based UAV Path Planning for 10-kV Distribution Networks Inspection in Live Working Scenarios

Dapeng Ma, Hongtao Jiang*, Lichao Jiang, Chi Zhang, Changwu Li, Xin Zheng, Mingxian Liu, Kai Li
Yunnan Power Grid Co., Ltd., Kunming, China

Abstract*—Before conducting maintenance on 10-kV distribution networks, the use of unmanned aerial vehicles (UAVs) for inspecting distribution lines can effectively enhance the operational efficiency of personnel in live working scenarios. For UAV-based inspection of power distribution networks, an optimal flight path ensures both operational safety and comprehensive image acquisition in live working scenarios. Therefore, this study proposes a UAV path planning algorithm and an insulator defect classification model based on YOLOv11, aiming to develop a UAV system for live power line detection. Firstly, a UAV path planning model is established to minimize the flight path length and maximize the image acquisition range, which also considers the safety distance constraints between UAVs and live power lines. On this basis, the optimization strategy of the particle swarm optimization (PSO) algorithm is introduced into the marine predictors algorithm (MPA), and a hybrid PSO-MPA algorithm is designed to improve the convergence accuracy of the MPA algorithm and solve the proposed UAV planning model. In addition, an insulator defect detection model has been developed to accurately identify the image information collected by UAVs. In order to improve the accuracy of the YOLOv11 model, the task-separation assignment (TSA) module was introduced into the YOLOv11 model, and a TSA-YOLOv11 model was designed. Experimental results demonstrate that the proposed PSO-MPA algorithm achieves superior convergence accuracy compared to five algorithms, including PSO. When the UAV flight step size is one meter, the PSO-MPA algorithm reduces the objective function value by an average of 49.62% relative to the other algorithms. Additionally, the TSA-YOLOv11 model attained an average accuracy of 96.87% for the insulator defect classification problem.

Keywords*—Marine predictors algorithm; YOLOv11; defect classification; UAV path planning; live power lines

I. Introduction

In live working scenarios, maintenance of the 10-kV distribution network by staff mainly faces challenges such as high safety risks, long inspection times, and blind spots in the field of vision. Unmanned aerial vehicles (UAVs) for inspecting 10-kV distribution networks can reduce the risk of direct exposure of personnel to high voltage environments, greatly improving safety. In addition, UAVs have multi-angle inspection capabilities, which can quickly obtain image data, significantly improving inspection efficiency and accuracy, and can timely detect hidden dangers that are difficult for humans to detect. UAV technology has the advantages of safety, high coverage, and strong defect detection capability in distribution network inspection tasks. Therefore, the power company is exploring the application of UAVs in the inspection tasks of distribution networks. Specifically, UAV-based distribution networks inspection technology achieves accurate diagnosis of insulator defects and wire breakage faults through visual sensors [1]. In addition, UAV-based distribution networks inspection is more in line with the goals of new power system construction in terms of economy and environmental protection. For the UAV-based distribution networks inspection system, the UAV path planning module is the foundation of intelligent inspection of distribution networks. This module automatically solves an optimal and safe flight route, aiming to shorten the inspection time or distance, avoid obstacles, and ensure flight safety, ultimately achieving comprehensive image acquisition throughout the inspection range. In addition, the insulator defect detection module aims to automatically, quickly, and accurately identify various defects in insulators from a large number of inspection images, thereby effectively reducing operation and maintenance costs, preventing power grid accidents, and improving power supply reliability [2]-[3].

At present, power companies require UAVs to operate in live power line scenarios. However, the current UAV path planning problem has not taken into account the safety distance constraints between UAVs and live power lines, and there are still problems such as low convergence accuracy of path optimization algorithms [4]. In addition, the accuracy of insulator defect identification is also an important challenge. Specifically, convolutional neural network (CNN) models face challenges such as complex background interference and high training data acquisition costs in insulator defect recognition tasks. The distribution networks background of the insulator captured by UAV is too complex, such as wires, towers, sky, and vegetation, which may affect the detection accuracy of the CNN model. Therefore, it is necessary to develop an insulator defect recognition model with strong feature extraction and classification capabilities to accurately locate and identify low contrast defects on insulators in background noise. In addition, CNN models require a large amount of high-quality and accurately labeled training data during the training process. Obtaining balanced data integration that covers various defect types, camera angles, and insulator models is costly. In the context of distribution networks in reality, there are far more normal insulator samples than defect samples, which leads to CNN-based insulator defect recognition models being prone to

*Corresponding author.

bias towards the "No Defects" category, significantly reducing the accuracy of defect recognition [5].

YOLOv8, as a real-time object detection model, is also used in insulator defect detection tasks. Although YOLOv8 performs better in small object detection tasks compared to CNN models, its detection accuracy is also lower for extremely fine cracks or damages on insulators, resulting in a higher rate of missed detection of surface defects on insulators. Overall, the biggest challenges for CNN and YOLOv8 in insulator defect recognition include complex background interference and low detection accuracy for small targets [6].

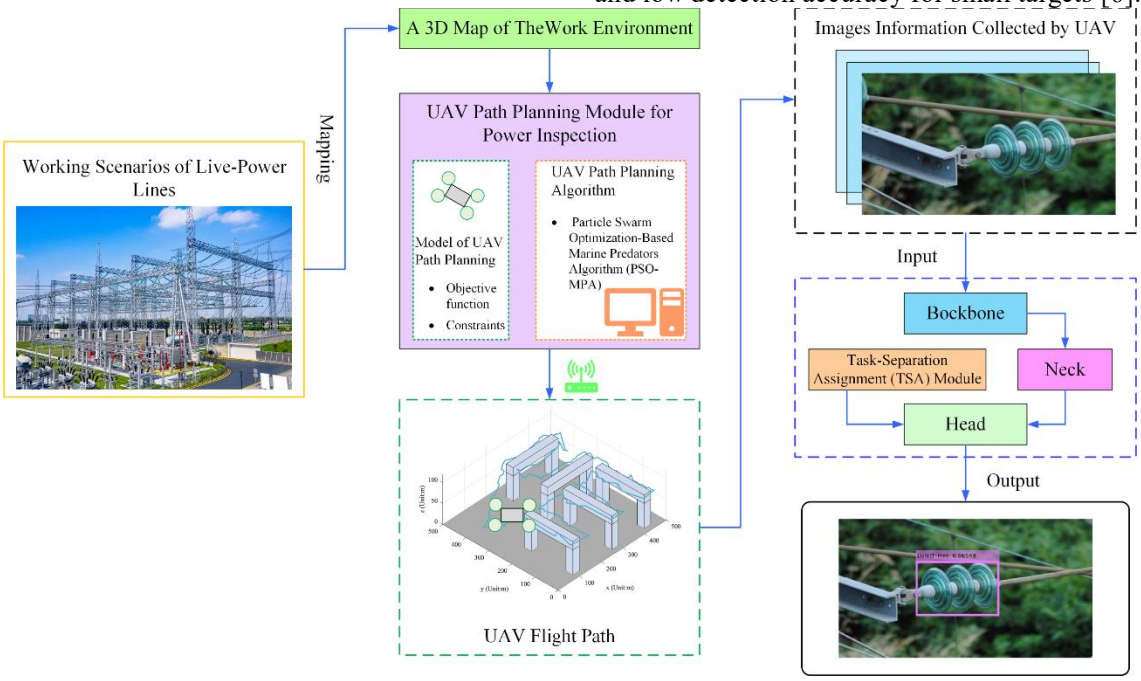


Fig. 1. UAV-based live power line inspection system.

The task paragraph assignment (TSA) module can further classify candidate samples based on specific defect types and confidence levels in insulator defect recognition tasks, thereby improving the accuracy of insulator defect recognition. Therefore, in this study, the TAS module was introduced into the head module of the YOLOv11 model with the aim of improving its accuracy. In addition, a biological heuristic algorithm has been developed to solve the optimal path for the UAV, aiming to safely and efficiently complete distribution networks detection tasks. Fig. 1 shows the UAV-based distribution networks detection system developed in this study. The system consists of two modules: UAV path planning and insulator defect recognition. The particle swarm optimization-based marine predictors algorithm (PSO-MAP) is used to optimize the flight path of inspection UAVs. Through this flight path, UAVs can collect a large amount of image data. Based on these image data, the insulator defect recognition module identifies and detects defects in insulators. The main contributions of this study are summarized as follows:

- A UAV-based intelligent detection system framework for live power lines has been designed. The system integrates a UAV path planning algorithm and an insulator defect recognition model, and develops a complete end-to-end UAV-based distribution network detection system.
- A mathematical model for path planning for UAV power line inspection scenarios has been developed, which aims to minimize the flight path length and maximize the image acquisition range. In addition, the model considers the safety distance constraint between UAVs and live lines to ensure operational safety.
- A hybrid PSO-MPA algorithm was developed by combining particle swarm optimization (PSO) and marine predator algorithm (MPA) optimization strategies. Experimental results have shown that this algorithm significantly improves the convergence accuracy of the PSO algorithm, and the potential of the PSO-MPA algorithm in optimizing the path of the inspection UAV has been demonstrated.
- The TSA module was introduced into the head module of the YOLOv11 model, and the TSA-YOLOv1 model was developed to improve the accuracy of the insulator defect recognition module based on the YOLOv11 model. The TSA-YOLOv11 model achieved an average accuracy of 96.87% in insulator defect classification tasks, improving the reliability of insulator defect detection.

The remaining parts of this study are arranged as follows: Section II reviews the work on UAV path planning and defect classification. Section III introduces the proposed UAV-based distribution networks inspection system. Section IV presents the results and discusses them. Finally, Section V summarizes the entire study.

II. Related Work

A. UAV Path Planning

At present, UAVs have been applied in various fields such as logistics, communication, and patrol. Therefore, a large number of UAV path planning models have been developed for these scenarios. In [7], the authors proposed an autonomous trajectory planning method for UAVs used in industrial facility inspections. This study designed a viewpoint resampling mechanism to meet the multi-perspective coverage requirements in complex industrial scenes. By dynamically optimizing the position of observation points, the image acquisition quality of key areas was improved, and the flight path length was shortened. Sonny et al. studied the path planning problem of UAV-assisted wireless networks and proposed an improved PSO algorithm [8]. In order to address the low convergence accuracy of the PSO algorithm in handling UAV path planning problems, this study introduced dynamic inertia weights and mutation operators into the PSO algorithm, aiming to enhance the algorithm's global search capability, minimize UAV energy consumption, and maximize ground user network coverage. Rückin et al. developed an active learning path planning framework for the semantic mapping of UAVs. This framework models path planning as an information gain maximization problem, evaluates the uncertainty of unexplored areas in real-time through Bayesian inference, dynamically guides UAVs to collect high-value data, and aims to plan a path for UAVs to explore unknown areas [9].

In [10], the authors propose a noise-aware path and energy management joint optimization model for collaborative operation of multiple UAVs in noise-sensitive areas. This study aims to quantify the cost of sound pollution by establishing a noise propagation model of the rotor, with the aim of constraining UAV flight altitude and path to avoid residential areas. At the same time, the study designed an intelligent power allocation strategy for batteries, aimed at maximizing the endurance of UAVs. In [11], the authors studied the challenges of UAV path planning under the framework of Internet of Drones (IoD), and proposed a distributed collaborative optimization framework. This study uses transmission delay cost and packet loss rate as path cost functions to generate paths for multiple UAVs.

In [12], the authors developed a distribution UAV autonomous path planning and control system based on the Internet of Things (IoT) and edge computing technology. The system deploys edge nodes to process UAV perception data in real-time and combines with a cloud-based task scheduling center to generate the global optimal path, significantly reducing the latency of end-to-end decision-making. The experimental results show that this path planning method can effectively reduce the communication cost and flight distance of UAVs. Yanmaz et al. developed a dynamic multi-objective path planning method to address the challenges of collaborative search and communication support for multiple UAVs. This method takes maximizing the target discovery probability and network connectivity as the objectives of the path planning model, establishes a Markov decision model, and designs a distributed UAV path optimization algorithm, aiming to dynamically adjust the UAV formation by real-time evaluation of communication link quality [13].

In addition to the PSO-based path optimization algorithm mentioned earlier, some other cutting-edge research is also exploring the application of biological heuristic algorithms in UAV path planning problems. For example, in [14], a UAV covert path planning method based on reinforcement learning (RL) was proposed. This method is inspired by the kinematics of natural organisms, and trains RL agents to learn strategies that mimic natural organisms' covert approach to prey by constructing a three-dimensional dynamic environment of state and action space. In [15], the authors developed a UAV path planning algorithm that combines Q-learning and Grey Wolf Optimizer (GWO) to solve complex 3D path planning problems for UAV flight. This study utilizes Q-learning to dynamically optimize key parameters of the GWO, such as convergence factor and population position update weights, in order to enhance the algorithm's adaptability in dynamic obstacle environments and improve the convergence accuracy of GWO algorithm.

B. Target Identification Model

In [16], the authors designed a YOLO architecture for identifying pedestrians and vehicles for target detection tasks aimed at border patrol. In addition, this study decomposes the complex collaborative patrol problem into global path planning and local conflict resolution and task allocation problems, aiming to efficiently coordinate the flight paths and monitoring tasks of multiple drones. Chen et al. focused on the challenging problem of detecting small ship targets in ocean-wide area remote sensing images. In order to overcome the difficulties of small targets having limited information and being easily submerged in the ocean background, this study also designed an improved YOLO model, aiming to enhance the model's ability to extract features and utilize motion information of small targets, thereby improving the accuracy of ship recognition [17]. In [18], the authors developed an insulator identification and defect detection system for distribution networks inspection. This study proposes a defect detection method for insulators based on YOLOX. The core of this method lies in optimizing the YOLOX model, introducing a high-resolution differential module aimed at capturing the local detailed features and minor defects of insulators and their key components more finely.

Zhang et al. proposed a method for identifying the working angle of a loader bucket based on the YOLOv5s model. This method utilizes YOLOv5s to quickly and accurately locate the target of the bucket, and introduces an attention mechanism to more accurately recognize the posture of the bucket, ultimately calculating the working angle of the bucket [19]. In [20], the authors focus on the task of detecting small targets in infrared images. In order to solve the problem of target loss or false alarm caused by traditional methods, this study proposes an improved strategy based on pixel extension. In [21], the authors developed an intelligent inspection system based on a UAV for power insulator inspection tasks under complex weather conditions. The system has successfully achieved autonomous positioning, state recognition, and defect detection of insulators by equipping sensors and machine learning algorithms suitable for complex environments. Xuan et al.

proposed an intelligent identification method for insulator defects based on Fully Convolutional One Stage (FCOS). The experimental results show that this method can not only accurately locate and segment individual insulators in the image, but also finely segment the defect areas on the insulator surface [22]. In [23], the authors proposed an efficient personal protective equipment detection model based on YOLOv8, aiming to develop a system for detecting whether workers are wearing compliant personal protective equipment, such as safety helmets, reflective clothing, and gloves, and to improve the accuracy and robustness of the YOLOv8 model in human protective equipment detection tasks.

III. METHODOLOGY

As shown in Fig. 1, the proposed UAV-based distribution networks inspection system consists of two parts: path planning and insulator detection. The UAV path planning model aims to minimize the flight path length and maximize the image acquisition range, while considering three constraints: the maximum pitch angle, maximum endurance time, and safe distance between the UAV and the live-line. On this basis, a PSO-MPA algorithm was developed for the path planning model of power inspection drones to solve the flight path of the drones. The TSA module was introduced into the YOLOv11 model to improve the accuracy of the insulator defect recognition model.

A. UAV Path Planning for Power Inspection

Before conducting UAV path planning, the transmission facilities are first divided into grids. In this study, the transmission facilities were surface divided into a set $V=\{v_1,v_2,,v_{b_\max}\}$. Based on the height of the transmission facilities, the initial flight height of the drone was determined as H_{fly}. The UAV's path consists of a series of path nodes $\Gamma=\{1,2,\cdots,m\}$. Another key parameter for UAV path planning is *Step_Size*.

The path distance cost f_{Path_Dis} of UAV is defined in Eq. (1):

$$f_{Path_Dis}=\sum_{m\in\Gamma}\sqrt{(x_m-x_{m-1})^2+(y_m-y_{m-1})^2+(z_m-z_{m-1})^2} \tag{1}$$

where, (x_m,y_m,z_m) is the position of the UAV's path node m in the coordinate system.

The definition of *Step_Size* is as follows in Eq. (2):

$$Step_Size=\sqrt{(x_m-x_{m-1})^2+(y_m-y_{m-1})^2+(z_m-z_{m-1})^2} \tag{2}$$

z_m is defined in Eq. (3):

$$z_m=H_{fly},\forall m\in\Gamma \tag{3}$$

The image acquisition range of the UAV is defined as Eq. (4):

$$f_{img}=\frac{\sum\beta(v_b)}{b_\max},\forall b\in\{1,2,\cdots,b_\max\} \tag{4}$$

where, $\beta(v_b)$ is a 0-1 variable. When the UAV captures an image of grid v_b position, $\beta(v_b)=0$; Otherwise, $\beta(v_b)=1$.

The objective function of the UAV path planning model used for distribution networks inspection is Eq. (5):

$$F=w_1\times f_{Path_Dis}+w_2\times f_{img} \tag{5}$$

where, w_1 and w_2 are the weight coefficients of the objective function.

The constraint conditions for UAV path planning in distribution networks inspection are defined as follows [Eq. (6)]:

$$\theta_m\le\theta_{\max},\forall m\in\Gamma \tag{6}$$

where, θ_m is the pitch angle of the distribution networks inspection UAV. $\theta_{\max}$ is the maximum pitch angle of the distribution networks inspection UAV [Eq. (7)].

$$\frac{f_{Path_Dis}}{speed_{UAV}}\le T_{UAV} \tag{7}$$

where, $speed_{UAV}$ is the flight speed of the distribution networks inspection UAV. T_{UAV} is the maximum endurance time of the distribution networks inspection UAV [Eq. (8)].

$$L(v_b,UAV)\ge L_{safety},\forall b\in\{1,2,\cdots,b_\max\} \tag{8}$$

where, $L(v_b,UAV)$ is the distance between the UAV and the distribution networks. L_{safety} is the safe distance between the UAV and the distribution networks.

To solve this path planning model, a PSO-MPA algorithm was developed. The MPA algorithm mainly includes predation strategies based on Brownian motion and target aggregation strategies based on Levy motion [24]. In order to improve the convergence accuracy of MPA, a speed update-based foraging strategy based on the PSO algorithm is also added to the hybrid PSO-MPA algorithm. The definition of the PSO-MPA algorithm is as follows:

1) Initialization: Initialize the position O of the marine predator population, initialize the speed of PSO, and set the maximum iteration number *Iter*_max of the PSO-MPA algorithm.

2) Calculate adaptive parameters: Determine the update and search process that should be executed for the current iteration count *Iter* based on the value of *Iter* / *Iter*_max.

3) Update strategy 1: The foraging strategy based on speed updates the positions of all marine predators to obtain O^{Δ}.

4) Update strategy 2: Based on Brownian motion, update the positions of some marine predators using a predation strategy. The update process is defined as Eq. (9):

$$O(k)_new=PBM(O(k)^{\Delta}) \tag{9}$$

where, $O(k)$ is defined as the location of the k-th marine predator. $PBM(.)$ is defined as a predation strategy based on Brownian motion.

5) Update strategy 3: Implement a target aggregation strategy based on Levy motion for the remaining marine predators.

6) Evaluate and select the optimal solution: Calculate the fitness of all updated positions of marine predators and generate the optimal solution.

B. Defect Detection of Insulators

Although multiple frameworks such as Real-Time Detection Transformer (RTDRTR), Region-based Convolutional Neural Networks (RCNN), and YOLOv8 have been used for specific object detection or insulator defect identification tasks, improving the accuracy of insulator defect detection models remains a challenge [23]. Therefore, in this study, the TSA module was introduced into the YOLOv11 model to improve the accuracy of insulator defect detection. Fig. 2 shows the header module of the proposed TSA-YOLOv11 model.

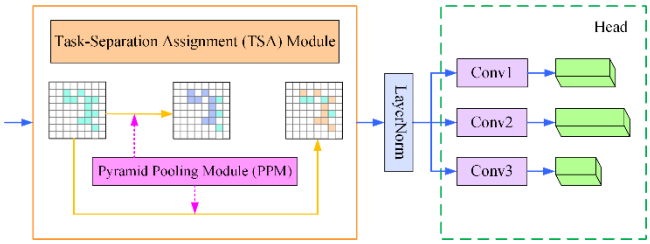


Fig. 2. Head module of the YOLOv11 model based on TSA (TSA-YOLOv11).

The header module of the TSA-YOLOv11 model is designed to enhance the recognition of insulator defect types by integrating the global modeling capability of the Transformer with the local feature extraction ability of CNN. After a series of convolutional layers and Bottleneck modules for dimensionality reduction and local feature extraction, the resulting insulator features are fed into the TSA module. Within this module, the features undergo multi-head integration, residual connections, and layer normalization before being output, thereby modeling dependency relationships between any two spatial positions.

In the header module of the TSA-YOLOv11 model, the pooling layer uses max pooling technology. The definition of max pooling is as follows [Eq. (10)]:

$$\mathrm{H}_j = Rand(0,1)\times \max\left\{\mathrm{A}_i, i\in P_j\right\} + (1-Rand(0,1))\times\frac{1}{|P_j|}\times\sum_{i\in P_j}\mathrm{A}_i \quad (10)$$

where, $Rand(0,1)$ is a random number with a value of 0 or 1. H_j is the j-th element of the feature map output from the max pooling operation. A_i is the i-th position of the j-th element in the input feature map P_j.

IV. Experiments

This study proposes an intelligent UAV inspection system for distribution networks detection, with a primary focus on addressing UAV path planning and insulator defect detection challenges. To validate the effectiveness of the proposed path planning algorithm, a three-dimensional flight simulation environment was established based on actual substation map data. For insulator defect detection, the image samples were collected under varying lighting conditions, shooting distances, and camera angles to ensure comprehensive coverage of operational scenarios.

A. Path Planning Results

In order to demonstrate the application results of the proposed PSO-MPA algorithm in the path planning problem of power inspection UAVs, and to explore the influence of the key parameter of flight step size on the objective function, multiple power inspection UAV path planning experiments were designed based on different flight step sizes. Firstly, the flight step sizes were set to 1 meter, 3 meters, 5 meters, 7 meters, 9 meters, and 11 meters, respectively. Based on different flight step sizes, independent experiments were conducted to explore the influence of flight step sizes on the fitness function. In addition, rime (RIME) optimization algorithm [25], Harris hawks optimization (HHO) [26], snow geese algorithm (SGA) [27], PSO, The Tornado optimizer with Coriolis force (TOC) algorithm is also used to solve the path planning problem of power inspection UAVs, aiming to verify the convergence accuracy and robustness of the PSO-MPA algorithm. In addition, in this study, standard deviation was used to evaluate the stability of the path planning algorithm for power inspection UAVs.

Fig. 3 shows the flight paths generated by the PSO-MPA algorithm for the power inspection UAV under different step sizes. Fig. 4(a) and Fig. 4(b) show the iterative function curves of PSO-MPA, HHO, PSO, RIME, SGA, and TOC algorithms with step sizes set to 1 meter and 3 meters, respectively. Fig. 5(a) and Fig. 5(b) show the iteration function curves of each algorithm when the step size is set to 5 meters and 7 meters, respectively. Fig. 6(a) and Fig. 6(b) show the iteration function curves of each algorithm when the step size is set to 9 meters and 11 meters, respectively. Table I shows the fitness function values of each algorithm at different step sizes.

From Fig. 4(a), it can be seen that the PSO-MPA algorithm generates a fitness function value of 3175.06 in solving the path planning problem of the power inspection UAV. Compared with the other five algorithms, the PSO-MPA algorithm has the smallest fitness function value, when the flight step size of the UAV is 1 meter. From Fig. 4(b), it can be seen that when the UAV's flight step size is 3 meters, the PSO-MPA algorithm also has the lowest fitness function value, which is 3276.60, an average decrease of 44.53% compared to the other five algorithms. From Fig. 5(a) and Fig. 5(b), it can be seen that when the UAV's flight step size is 5 meters and 7 meters, respectively, compared with HHO, PSO, RIME, SGA, and TOC algorithms, the fitness function value of PSO-MPA algorithm is also the smallest. From Fig. 6(a) and Fig. 6(b), it can be seen that when the UAV's flight step size is 9 meters

and 11 meters, the fitness function values of the PSO-MPA algorithm are 3559.49 and 3232.06, respectively.

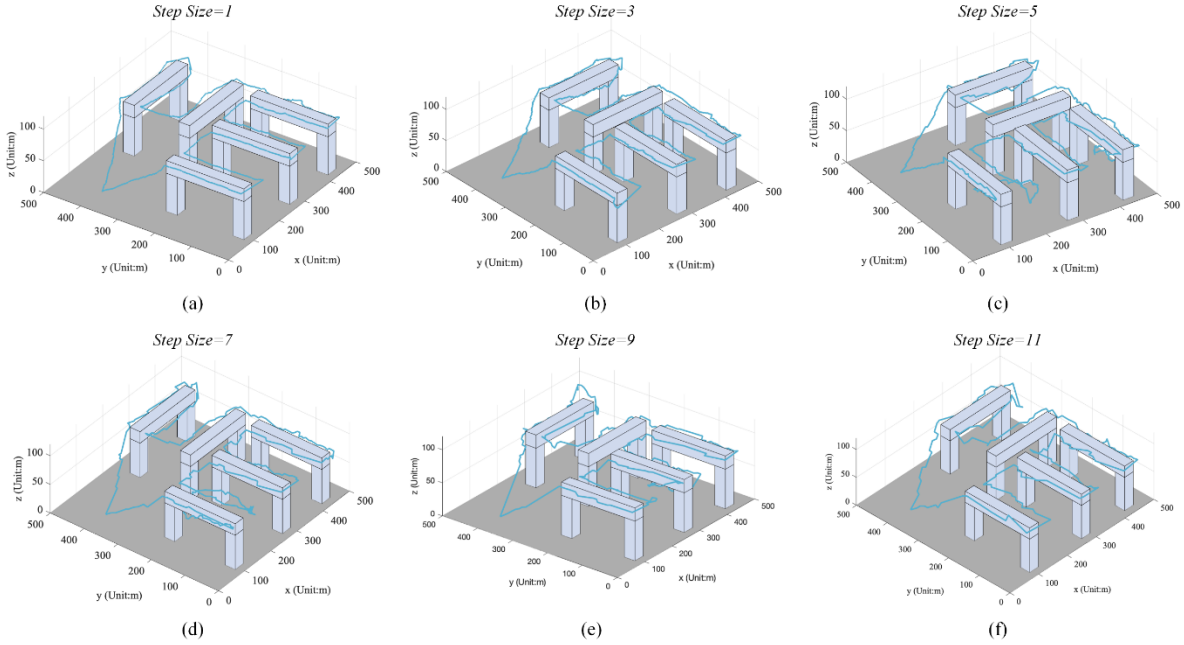


Fig. 3. UAV path planning results with different flight step sizes.

TABLE I. THE FITNESS FUNCTION VALUES OF UAV PATHS WITH DIFFERENT FLIGHT STEP SIZES

Algorithm	PSO-MPA	HHO	PSO	RIME	SGA	TOC
Step Size	*Fitness Function*					
1	3175.06	6665.64	6328.04	7242.72	8539.11	4343.94
3	3276.60	7626.56	6115.06	5133.10	7717.98	4396.74
5	3790.45	10572.84	7185.90	5184.23	7727.14	5169.95
7	3909.39	7121.22	6995.10	7638.17	7819.61	7842.44
9	3559.49	7914.00	7252.17	6775.45	8759.74	7146.64
11	3232.06	7058.16	5329.62	6135.50	5515.12	5837.17

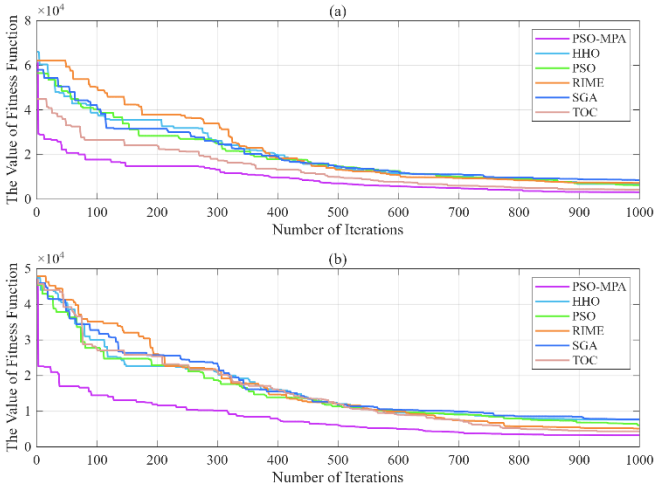


Fig. 4. The iterative function curve for flight steps of 1 meter and 3 meters.

Therefore, it can be concluded that among all step settings, the proposed PSO-MPA algorithm generates the smallest fitness function value compared to other algorithms when solving the path planning problem of power inspection UAV. The PSO-MPA algorithm performs significantly better than HHO, PSO, RIME, SGA, and TOC algorithms in solving path planning problems for power inspection UAVs. The objective function value corresponding to the PSO-MPA algorithm decreased by 25.48% to 47.46% compared to the objective function value generated by the suboptimal algorithm. When the step size is set to 9 meters, the fitness function value of the PSO-MPA algorithm is 3559.49, and the decrease in the objective function value is the most significant. Compared with the second ranked RIME algorithm, the objective function value of the PSO-MPA algorithm has decreased by 47.46%.

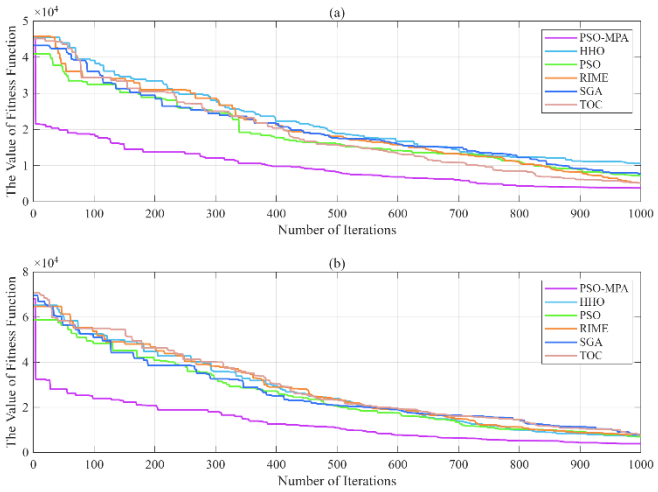


Fig. 5. The iterative function curve for flight steps of 5 meters and 7 meters.

When the step size is set to 3 meters, the performance of the PSO-MPA algorithm decreases, but compared with HHO, PSO, RIME, SGA, and TOC algorithms, the PSO-MPA algorithm still maintains a significant advantage in solving the path planning problem of power inspection UAVs. Compared with the second ranked TOC algorithm, the fitness function value of the PSO-MPA algorithm is reduced by 25.48%. In addition, by comparing the fitness values of each step size setting corresponding to the PSO-MPA algorithm, a flight step size of 1 meter was determined to be the optimal setting, and the generated drone path cost was the lowest.

Table II shows the standard deviation of the fitness function values generated by each algorithm running independently ten times at different step sizes. Fig. 7 shows the standard deviation of the objective function values generated by running different algorithms ten times with different step sizes. It can be seen that compared with HHO, PSO, RIME, SGA, and TOC algorithms, PSO-MPA is the overall most stable algorithm. The PSO-MPA algorithm achieved the minimum standard deviation in all five step settings, indicating that it can maintain high consistency under different parameters. Especially when step size is 5 meters, the standard deviation of the PSO-MPA algorithm reaches 95.93, indicating the highest stability. HHO is the most unstable algorithm overall, with a standard deviation exceeding 270 under multiple flight step settings, significantly higher than other algorithms.

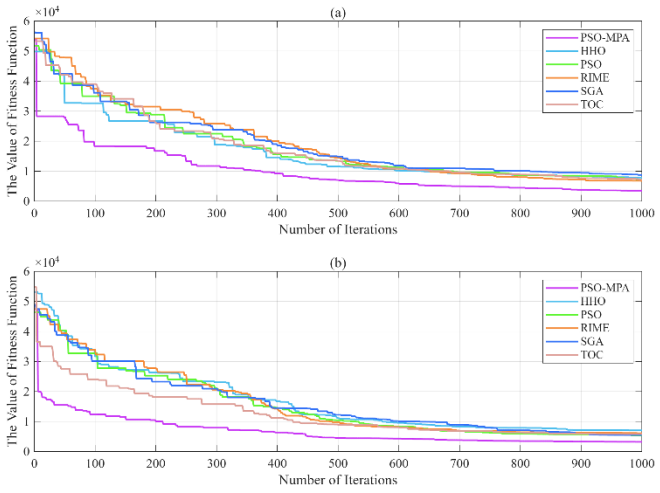


Fig. 6. The iterative function curve for flight steps of 9 meters and 11 meters.

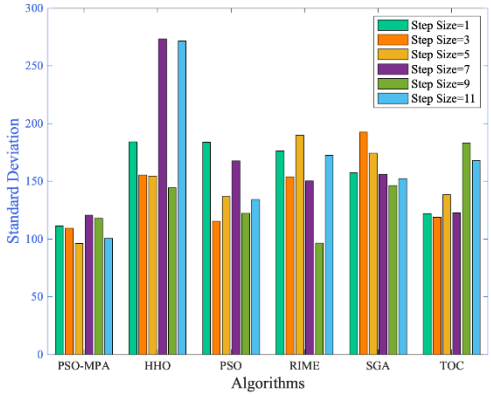


Fig. 7. The standard deviation of the fitness function values for UAV paths with different flight step sizes.

TABLE II. THE STANDARD DEVIATION OF THE FITNESS FUNCTION VALUES FOR UAV PATHS WITH DIFFERENT FLIGHT STEP SIZES

Step Size	Step Size=1	Step Size=3	Step Size=5	Step Size=7	Step Size=9	Step Size=11
Algorithm	***Standard Deviation***					
PSO-MPA	**111.25**	**109.27**	**95.93**	**120.58**	117.73	**100.46**
HHO	184.11	155.46	154.54	273.41	144.24	271.39
PSO	183.61	115.22	136.78	167.78	122.32	134.26
RIME	176.18	153.63	189.80	150.17	**96.32**	172.27
SGA	157.26	192.48	174.19	156.14	146.22	152.15
TOC	121.82	118.70	138.26	122.44	183.05	167.97

B. Defect Detection Results of Insulators

To verify the performance of the TAS-YOLOv11 model in insulator defect detection, insulator defect detection experiments were conducted based on the RTDETR model, R-CNN model, YOLOv8 model, YOLOv10 model, and YOLOv11 model [28]-[30]. Table III shows the types of insulator defects included in the training and testing sets of this study. In this study, insulators used in distribution networks were classified into four types: defect free, defect type 1, defect type 2, and defect type 3.

In this study, the dataset includes 8000 insulator images. The number of images in each of the four categories (insulators without defects, defect type 1, defect type 2, and defect type 3) is 2000. The training set includes 1600 insulator images for each of the four categories. The test set includes 1000 insulator images for each of the four categories. Table IV shows the parameter settings for different models. For different insulator defect detection models, the batch size and number of epochs are 2 and 300, respectively.

TABLE III. DEFECT TYPES OF INSULATORS

Class	Definition
Defect-Free	No defects: The insulator has no defects and works normally.
Defect Type 1	Surface Defect: The material on the surface of the insulator is damaged.
Defect Type 2	Damaged: The edge of the insulator is damaged or there are cracks in the insulator.
Defect Type 3	Missing: The insulator string is missing one or several insulators.

Table V presents the mean average precision (mAP@0.5) and training loss values for different insulator defect detection models. The proposed TAS-YOLOv11 achieves an mAP@0.5 of 95.16%, outperforming the RTDETR and R-CNN models by 11.08% and 6.43%, respectively. Additionally, its training loss is 0.1252, representing reductions of 37.59% and 36.28% compared to RTDETR and R-CNN, respectively. Compared with the RTDETR model, R-CNN model, YOLOv6 model, YOLOv8 model, YOLOv10 model, and YOLOv11 model, the mAP@0.5 index of TAS-YOLOv11 increased by an average of 5.14%, and the training loss of TAS-YOLOv11 decreased by an average of 28.49%. Fig. 8 shows the results of insulator defect recognition based on the TAS-YOLOv11 model.

TABLE IV. PARAMETER SETTINGS

Models	Parameter	Value
RTDETR	Batch size	2
	Number of epochs	300
	Learning rate	1E-4
R-CNN	Batch size	2
	Number of epochs	300
	Learning rate	1E-6
YOLO	Batch size	2
	Number of epochs	300
	Learning rate	1E-3

Table VI shows the performance of different insulator defect detection models. In Table VI, two indicators, accuracy and recall, are used to evaluate different insulator defect detection models. From Table VI, it can be seen that in the experiment of insulator defect detection, the TAS-YOLOv11 model has the best performance and highest accuracy in the four types of defect detection.

TABLE V. PERFORMANCE COMPARISON OF DIFFERENT MODELS

Models	mAP@0.5 (%)	Training Loss
RTDETR	85.67	0.2006
R-CNN	89.41	0.1965
YOLOv8	91.32	0.1887
YOLOv10	92.18	0.1609
YOLOv11	94.46	0.1435
TAS-YOLOv11	95.16	0.1252

TABLE VI. IDENTIFICATION RESULTS OF INSULATOR DEFECTS USING DIFFERENT MODELS

Models	Class	Accuracy	Recall
RTDETR	Defect-Free	0.8745	0.8617
	Defect Type 1	0.8659	0.8569
	Defect Type 2	0.8943	0.8722
	Defect Type 3	0.8741	0.8632
R-CNN	Defect-Free	0.8918	0.8774
	Defect Type 1	0.8901	0.8749
	Defect Type 2	0.8792	0.8564
	Defect Type 3	0.8577	0.8513
YOLOv8	Defect-Free	0.9674	0.9519
	Defect Type 1	0.9357	0.9349
	Defect Type 2	0.9290	0.9125
	Defect Type 3	0.9293	0.9123
YOLOv10	Defect-Free	0.9666	0.9667
	Defect Type 1	0.9381	0.9305
	Defect Type 2	0.9374	0.9181
	Defect Type 3	0.9341	0.9183
YOLOv11	Defect-Free	0.9683	0.9557
	Defect Type 1	0.9375	0.9277
	Defect Type 2	0.9434	0.9320
	Defect Type 3	0.9326	0.9379
TAS-YOLOv11	Defect-Free	0.9793	0.9632
	Defect Type 1	0.9653	0.9564
	Defect Type 2	0.9699	0.9417
	Defect Type 3	0.9603	0.9581

The defect free insulator detection accuracy and recall of TAS-YOLOv11 model are 97.93% and 96.32%, respectively. Compared to the accuracy of the YOLOv11 model, the TAS-YOLOv11 model shows an improvement of over 1 percentage point. In the detection of the most difficult surface defect (defect type 1), the accuracy of the TAS-YOLOv11 model is 96.53%, which is 2.72 percentage points higher than the suboptimal model YOLOv10. The recall rate of traditional R-

CNN models in missing defects (defect type 3) is only 85.13%, which poses a high risk of inaccurate recognition.

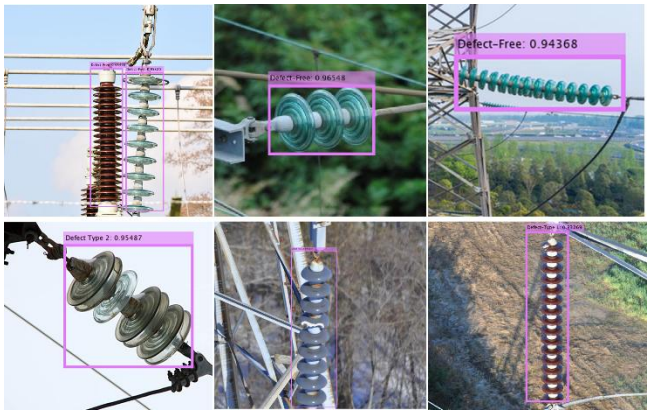


Fig. 8. The results of insulator defect detection based on TAS-YOLOv11 model.

V. Conclusion

This study proposes a dual-layer detection model for distribution networks, integrating UAV path planning and insulator defect recognition. In the path planning layer, a power grid inspection UAV path planning model aims to minimize flight distance and maximize image coverage. In addition, a hybrid PSO-MPA algorithm is designed to solve the path planning model of UAV for power grid inspection. The experimental results demonstrate that with a flight step size of one meter, the PSO-MPA algorithm reduces the objective function value by an average of 49.62% compared to the other five algorithms, significantly shortening the path length and improving the integrity of image acquisition for transmission facilities. In the insulator defect recognition layer, a TSA-YOLOv11 model was developed, which introduces a task decoupling allocation module (TSA) to separate positioning and classification tasks, aiming to solve the feature conflict problem of traditional models. The average accuracy of insulator defect detection reached 96.87%, and the recall rate of high-risk damaged defects increased to 94.17%. In future research, combining reinforcement learning with proposed models to construct an online inspection system will be a key focus.

Acknowledgment

This work was supported by the Science and Technology Project of China Southern Power Grid Co., Ltd. under Grants YNKJXM20240219 and YNKJXM20240472.

References

[1] H. Liu et al., "Study on UAV Parallel Planning System for Distribution networks Project Acceptance Under the Background of Industry 5.0," IEEE Transactions on Industrial Informatics, vol. 18, no. 8, pp. 5537-5546, Aug. 2022, doi: 10.1109/TII.2022.3142723.

[2] P. Hamelin et al., "Discrete-time control of LineDrone: An assisted tracking and landing UAV for live power line inspection and maintenance," 2019 International Conference on Unmanned Aircraft Systems (ICUAS), Atlanta, GA, USA, 2019, pp. 292-298, doi: 10.1109/ICUAS.2019.8798137.

[3] N. Jain, J. Bedi, A. Anand and S. Godara, "A Transfer Learning Architecture to Detect Faulty Insulators in Powerlines," IEEE Transactions on Power Delivery, vol. 39, no. 2, pp. 1002-1011, April 2024, doi: 10.1109/TPWRD.2024.3353203.

[4] Haizhou Zhang and Shengnan Xu, "Path Planning Technology for Unmanned Aerial Vehicle Swarm Based on Improved Jump Point Algorithm" International Journal of Advanced Computer Science and Applications (IJACSA), 16 (4), 2025. http://dx.doi.org/10.14569/IJACSA.2025.0160426

[5] Wang Tingyu, Sun Xia, Liu Jiaxing and Zhang Yue, "A Deep Learning Based Detection Method for Insulator Defects in High Voltage Distribution networks" International Journal of Advanced Computer Science and Applications (IJACSA), 15 (10), 2024. http://dx.doi.org/10.14569/IJACSA.2024.0151040

[6] M. He, L. Qin, X. Deng and K. Liu, "MFI-YOLO: Multi-Fault Insulator Detection Based on an Improved YOLOv8," IEEE Transactions on Power Delivery, vol. 39, no. 1, pp. 168-179, Feb. 2024, doi: 10.1109/TPWRD.2023.3328178.

[7] H. Liu, Y. P. Tsang, C. K. M. Lee and C. H. Wu, "UAV Trajectory Planning via Viewpoint Resampling for Autonomous Remote Inspection of Industrial Facilities," IEEE Transactions on Industrial Informatics, vol. 20, no. 5, pp. 7492-7501, May 2024, doi: 10.1109/TII.2024.3361019.

[8] A. Sonny, S. R. Yeduri and L. R. Cenkeramaddi, "Autonomous UAV Path Planning Using Modified PSO for UAV-Assisted Wireless Networks," IEEE Access, vol. 11, pp. 70353-70367, 2023, doi: 10.1109/ACCESS.2023.3293203.

[9] J. Rückin, F. Magistri, C. Stachniss and M. Popović, "An Infomative Path Planning Framework for Active Learning in UAV-Based Semantic Mapping," IEEE Transactions on Robotics, vol. 39, no. 6, pp. 4279-4296, Dec. 2023, doi: 10.1109/TRO.2023.3313811.

[10] D. Scott, S. G. Manyam, I. E. Weintraub, D. W. Casbeer and M. Kumar, "Noise Aware Path Planning and Power Management of Hybrid Fuel UAVs," IEEE Transactions on Automation Science and Engineering, vol. 22, pp. 8227-8238, 2025, doi: 10.1109/TASE.2024.3481998.

[11] J. V. Shirabayashi and L. B. Ruiz, "Toward UAV Path Planning Problem Optimization Considering the Internet of Drones," IEEE Access, vol. 11, pp. 136825-136854, 2023, doi: 10.1109/ACCESS.2023.3339227.

[12] Xiuzhu Zhang, "Path Planning and Control of Intelligent Delivery UAV Based on Internet of Things and Edge Computing" International Journal of Advanced Computer Science and Applications (IJACSA), 15 (3), 2024. http://dx.doi.org/10.14569/IJACSA.2024.01503107

[13] E. Yanmaz, H. M. Balanji and İ. Güven, "Dynamic Multi-UAV Path Planning for Multi-Target Search and Connectivity," IEEE Transactions on Vehicular Technology, vol. 73, no. 7, pp. 10516-10528, July 2024, doi: 10.1109/TVT.2024.3363840.

[14] J. Li, Y. Zhu, C. Li and Z. Song, "A Motion Camouflage-Inspired Path Planning Method for UAVs Based on Reinforcement Learning," IEEE Transactions on Aerospace and Electronic Systems, vol. 61, no. 2, pp. 4105-4114, April 2025, doi: 10.1109/TAES.2024.3496417.

[15] Binbin Tu, Fei Wang, Xiaowei Han and Xibei Fu, "Q-learning Guided Grey Wolf Optimizer for UAV 3D Path Planning" International Journal of Advanced Computer Science and Applications (IJACSA), 15 (7), 2024. http://dx.doi.org/10.14569/IJACSA.2024.0150747

[16] Haishi Liu, Yuxuan Sun, Nan Pan, Qiyong Chen, Xiaojue Guo, Dilin Pan, "Multi-UAV Cooperative Task Planning for Border Patrol based on Hierarchical Optimization" in Journal of Imaging Science and Technology, vol.65, pp 040402-1 - 040402-8, 2021, doi: https://doi.org/10.2352/J.ImagingSci.Technol.2021.65.4.040402

[17] J. Chen, Z. Hu, W. Wu, Y. Zhao and B. Huang, "LKPF-YOLO: A Small Target Ship Detection Method for Marine Wide-Area Remote Sensing Images," IEEE Transactions on Aerospace and Electronic Systems, vol. 61, no. 2, pp. 2769-2783, April 2025, doi: 10.1109/TAES.2024.3476459.

[18] Y. Li, D. Feng, Q. Zhang and S. Li, "HRD-YOLOX Based Insulator Identification and Defect Detection Method for Distribution networks," IEEE Access, vol. 12, pp. 22649-22661, 2024, doi: 10.1109/ACCESS.2024.3363430.

[19] X. Zhang, B. Cui, Z. Wang and W. Zeng, "Loader Bucket Working Angle Identification Method Based on YOLOv5s and EMA Attention

Mechanism," IEEE Access, vol. 12, pp. 105488-105496, 2024, doi: 10.1109/ACCESS.2024.3435146.

[20] Z. Zhao, H. Wang, H. Li, J. Yang and P. Yu, "A Pixel Expansion-Based Improvement in Dense Nesting Structures for Infrared Small Target Detection," IEEE Geoscience and Remote Sensing Letters, vol. 22, pp. 1-5, 2025, Art no. 6005205, doi: 10.1109/LGRS.2025.3547899.

[21] Y. Li, Y. Lu, K. Wu, Y. Fang, C. Zheng and J. Zhang, "Intelligent Inspection System for Power Insulators Based on AAV on Complex Weather Conditions," IEEE Transactions on Applied Superconductivity, vol. 34, no. 8, pp. 1-4, Nov. 2024, Art no. 9003204, doi: 10.1109/TASC.2024.3465368.

[22] Z. Xuan, J. Ding and J. Mao, "Intelligent Identification Method of Insulator Defects Based on CenterMask," IEEE Access, vol. 10, pp. 59772-59781, 2022, doi: 10.1109/ACCESS.2022.3179975.

[23] Z. Wang, Y. Zhu, Z. Ji, S. Liu and Y. Zhang, "An Efficient YOLOv8-Based Model With Cross-Level Path Aggregation Enabling Personal Protective Equipment Detection," IEEE Transactions on Industrial Informatics, vol. 20, no. 11, pp. 13003-13014, Nov. 2024, doi: 10.1109/TII.2024.3431045.

[24] A. Faramarzi, M. Heidarinejad, S. Mirjalili, and A. H. Gandomi, "Marine Predators Algorithm: A nature-inspired metaheuristic," Expert Systems With Applications, vol. 152, Art. no. 113377, 2020, doi: 10.1016/j.eswa.2020.113377

[25] H. Su et al., "RIME: A physics-based optimization," Neurocomputing, vol. 532, pp. 183–214, 2023, doi: 10.1016/j.neucom.2023.02.010

[26] 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, doi: 10.1016/j.future.2019.02.028

[27] A.-Q. Tian, F.-F. Liu, and H.-X. Lv, "Snow Geese Algorithm: A novel migration-inspired meta-heuristic algorithm for constrained engineering optimization problems," Applied Mathematical Modelling, vol. 126, pp. 327–347, 2024, doi: 10.1016/j.apm.2023.10.045

[28] Sun, H., Wang, Y., Du, J., & Wang, R. (2025). MFE-YOLO: A Multi-feature Fusion Algorithm for Airport Bird Detection. ICCK Transactions on Intelligent Systematics, 2(2), 85–94. https://doi.org/10.62762/TIS.2025.323887

[29] Hamza, M., Ali, I., Ali, S., Khan, W., Shah, S.M., & Haq, I.U. (2025). Leveraging Machine Learning and Deep Learning for Advanced Malaria Detection Through Blood Cell Images. ICCK Journal of Image Analysis and Processing, 1(1), 17–26. https://doi.org/10.62762/JIAP.2025.514726

[30] Akhtar, F., & Mahum, R. (2025). IRV2-hardswish Framework: A Deep Learning Approach for Deepfakes Detection and Classification. ICCK Journal of Image Analysis and Processing, 1(1), 45–56. https://doi.org/10.62762/JIAP.2025.421251

Enhancing Predictive Maintenance Method Using Machine Learning to Improve IoT-Embedded Machinery Efficiency and Performance

Abiinesh Nadarajan, Iskandar Ishak, Noridayu Manshor, Raihani Mohamed, Mohamad Yusnisyahmi Yusof
Department of Computer Science-Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia

***Abstract*—Predictive maintenance plays a crucial role in minimizing unplanned downtimes, reducing maintenance costs, and optimizing the operational efficiency of IoT-embedded industrial machinery. Despite its transformative potential, traditional predictive maintenance methods often face challenges such as limited accuracy, high latency, and inefficiencies in processing large and imbalanced datasets. This study proposes an enhanced predictive maintenance method using the Sliding Window Method with XGB model (E.XGB), incorporating advanced data preprocessing, permutation importance, and hyperparameter optimization to address these limitations. The proposed method was evaluated on two datasets, which are the synthetic AI4I 2020 Predictive Maintenance Dataset and the real-world CNC Milling Dataset. A comparative analysis with a predictive maintenance method using E.AB from prior research as a benchmark, along with several baseline models, DT, RF, and SVM, revealed that the E.XGB model consistently outperformed other methods in accuracy, precision, recall, and F1-scores. On the AI4I2020 dataset, the E.XGB model achieved an accuracy of 99.05%, while on the CNC Milling dataset, it attained an accuracy of 99.01%. Additionally, the E.XGB model also demonstrated reduced training and prediction times, meeting the real-time requirements of industrial applications. The proposed model demonstrated training speed of approximately 94% and prediction speeds of approximately 99.8% improvement over the E.AB model, making it highly suitable for real-time industrial applications. By improving accuracy, training speed, and prediction latency, the predictive maintenance method offers a robust, scalable, and reliable solution for predictive maintenance across diverse industrial contexts.**

Keywords*—Internet of Things; machine learning; predictive maintenance

I. Introduction

Industrial maintenance has evolved from reactive strategies to predictive maintenance, a proactive approach utilizing machine learning and IoT technologies. Traditional methods are often costly, inefficient, and prone to unplanned downtimes due to manual feature extraction and inaccurate degradation modelling [1, 2]. Predictive maintenance, however, predicts potential failures and optimizes schedules, allowing for timely interventions without detailed knowledge of degradation processes [3]. This approach offers significant benefits, such as reduced downtime, extended equipment life, and enhanced sustainability [4]. However, implementing predictive maintenance remains challenging due to the complexity of industrial systems and the intricate data patterns requiring analysis [5]. These challenges highlight the need for advanced methods that enhance accuracy and processing efficiency. By leveraging IoT and machine learning technologies, this study addresses these limitations, focusing on adaptability and scalability to meet the demands of diverse industrial applications.

The main problem that motivates this research is that current predictive maintenance methods still face challenges in achieving fast training speeds and prediction speeds while maintaining high accuracy, which is essential for real-time industrial applications [1, 6]. Due to massive sensor data generated by IoT-embedded machines, many existing methods rely on complex machine learning models to predict faults. Although this is effective and accurate, it leads to prolonged training times and higher prediction [7]. Having long training time can lead to delayed model updates, causing adaptability and scalability issues, while long prediction time causes delayed fault detection, which is important for real-time industrial machines. This study addresses these limitations by proposing an Enhanced Predictive Maintenance Method using the Sliding Window Method with XGB model (E.XGB) on sensor data, which can optimize training speeds and lower prediction latency while still maintaining a high accuracy [5]. By addressing these gaps, the study aims to enhance current predictive maintenance methods to satisfy the real-time requirements of industrial machinery.

The goal is to improve fault prediction accuracy, minimize false alarms, and reduce unnecessary maintenance and downtime. Faster training speeds and real-time predictions are also prioritized to support scalability and timely decision-making in industrial operations. The research focuses on IoT-driven machinery, using the AI4I 2020 Predictive Maintenance Dataset [8] and CNC Milling Dataset [9]. The proposed E.XGB model is evaluated against other machine learning methods like AdaBoost, Decision Tree, Random Forest, and Support Vector Machine. Performance is measured using metrics like accuracy, precision, recall, F1 score, ROC AUC, and prediction latency, with results benchmarked against existing approaches. This study's significance lies in improving fault detection accuracy, reducing training times, and enabling real-time predictions, which together lower downtime and maintenance costs. By bridging the gap between machine learning advancements and practical industrial applications, the research provides a robust foundation for future predictive maintenance innovations.

The rest of this study is organized as follows: Section II reviews related work on predictive maintenance, including algorithms, and its performance for IoT-based industrial systems. Section III details the methodology, including data preparation, model training, and optimization techniques. Section IV explains the proposed E.XGB method, supported by flowcharts, formulas, and calculations. Section V presents experimental results, comparing the proposed model's performance to previous methods. Section VI summarizes the study's contributions and provides suggestions for future research.

II. Related Works

In recent years, predictive maintenance for IoT-embedded industrial machinery has garnered significant attention. In [1], the authors proposed a predictive maintenance method that utilizes machine learning algorithms, particularly AdaBoost, to classify machine stops in knitting machines. Their system achieved 92% accuracy, enabling timely maintenance and improved efficiency in the textile industry. In terms of hyperparameter tuning, they employed grid search cross-validation (GridSearchCV) for hyperparameter optimization. However, the resulting output possessed high latency in processing and may hinder its application in real-time systems.

The challenge of imbalanced datasets is tackled by enhancing the KNN algorithm [10]. By employing techniques like feature engineering, standardization, and under-sampling, their model achieved a 97.1% accuracy. The ability to handle imbalanced data is crucial in predictive maintenance, although KNN's performance can vary based on dataset characteristics. The approach also explicitly mentions hyperparameter tuning and states that Grid Search tuning was applied to optimize the K-Nearest Neighbours (KNN) model. However, it does use an automated hyperparameter tuning method.

Similarly, in [11], the authors proposed a framework combining XGB and LOF models with XAI to address imbalanced datasets. Their method achieved 96% accuracy with XGB and 91% with LOF. The dual approach helps by focusing on predictive classification with XGB and anomaly detection with LOF. However, the study highlighted potential issues with the synthetic noise introduced by SMOTE and scalability challenges in real-time applications. However, no explicit mention of hyperparameter tuning was found in this study.

Multiple machine learning models including LR, KNN, and ANN, for failure prediction [12]. The ANN model outperformed others with an accuracy of 96.85%, but it struggled with false positives, indicating a need to balance sensitivity and specificity in complex environments. In this approach, structured hyperparameter tuning methods that include GridSearchCV and RandomizedSearchCV were used in their work.

LSTM networks combined with Auto Encoder is employed to predict tool wear in CNC milling machines [13]. Their model achieved 98% accuracy, showing improved performance over previous methods. However, long datasets could reduce the model's accuracy, and the authors recommended prioritizing relevant data through a weighting mechanism. Also, there was no hyperparameter tuning approach mentioned in the approach.

In another deep learning approach, reinforcement learning techniques like Q-learning and SARSA alongside LSTM networks is applied to monitor CNC tool conditions [14]. The SARSA algorithm achieved 98.66% accuracy. While reinforcement learning optimizes decision-making, the complexity of the model posed challenges in training, requiring large datasets and still misclassifying certain instances. Again, in this approach, no mention of hyperparameter tuning is involved.

Machine learning and deep learning methods for tool wear prediction were compared, showing that CNN and AE-LSTM outperformed KNN [15]. While deep learning models are more accurate, they are computationally intensive, raising concerns about their scalability. In this approach, there was no mention about the usage of hyperparameter tuning.

The use of SVM for detecting machine failures in predictive maintenance is compared in [16], achieving up to 88% accuracy in testing data. In terms of hyperparameter tuning, this approach employs grid search for selecting optimal hyperparameters (C and Gamma). However, the SVM's reliance on parameter tuning, such as C and gamma, can slow optimization, making it unsuitable for real-time applications.

LSTM models are used for predictive degradation modelling in milling machines, achieving 80% accuracy [17]. The advantage of this method lies in its ability to predict degradation by considering sequential dependencies. However, it struggled with imbalanced data and may not account for all failure modes, especially in more complex systems. On real-time systems, [2] integrated IoT and machine learning for predictive maintenance, using algorithms like SVM, RNN, and CNN. Their method achieved 87% accuracy with SVM on a milling dataset and 98% with CNN on a bearing dataset. While the models performed well, processing large datasets and extracting meaningful features for real-time predictive maintenance posed challenges.

A hybrid deep learning model combining CNN and LSTM for fault detection is proposed in [18], achieving 97.9% accuracy. The model's ability to capture both spatial and temporal patterns in multivariate sensor data made it effective for fault detection. However, its complexity and need for substantial annotated data slow down its real-time applicability.

A PM-C-LSTM model is used for wind turbines, achieving a 96.77% accuracy [19]. While promising for wind energy applications, the model's effectiveness across different environments requires further validation. Despite the advancements, the reviewed works reveal several gaps in predictive maintenance research. A significant issue is the trade-off between accuracy and processing speed. Deep learning models, like CNN-LSTM and hybrid models, achieve high accuracy but come with high computational costs, which may limit their use in real-time applications.

Models like AdaBoost [1] and SVM [16] also face challenges in terms of processing speed, as they require substantial parameter tuning, slowing down optimization. Managing imbalanced datasets remains another challenge.

While techniques like SMOTE and under-sampling improve model performance [10, 11], they may introduce synthetic noise or fail to generalize effectively. Furthermore, models may perform poorly when failure events are rare, highlighting the need for better methods to manage imbalanced data without sacrificing accuracy.

In summary, the research on predictive maintenance for IoT-embedded industrial machinery has made significant strides, but challenges in accuracy, processing speed, and handling imbalanced data persist. Moving forward, research should focus on optimizing the balance between accuracy and speed, improving the generalization of models across diverse datasets, and finding better methods to handle imbalanced data while avoiding the introduction of noise. Addressing these issues will pave the way for more efficient, dependable, and scalable predictive maintenance systems that can meet the demands of real-world industrial environments.

III. Methodology

A. Experimental Setup

This study evaluates the proposed Enhanced XGBoost (E.XGB) model for predictive maintenance by comparing it with the method by [1] and several baseline models. The E.XGB model builds upon the approach in [1] for IoT industrial applications, improving against the AB model with the proposed XGB model by incorporating minor enhancements. The method proposed by [1] is replicated in this study using the same configurations as the original study to be used as a comparison. The replicated Elkateb method, referred to as Enhanced AdaBoost (E.AB), is used as the primary benchmark. Additionally, three baseline models, DT, RF, and SVM, will serve as secondary benchmarks, using default parameters without optimization. All models will be trained and tested on both imbalanced and balanced datasets to ensure comprehensive evaluation. Performance is measured using metrics such as accuracy, precision, recall, F1-score, confusion matrices, ROC-AUC, training times, and prediction times. This study aims to address the challenges in the method in [1] and demonstrate the effectiveness of the proposed E.XGB model.

B. System Infrastructure

The predictive maintenance system for IoT-embedded machinery integrates real-time data collection, cloud-based processing, and machine learning to predict and prevent equipment failures. Fig. 1 illustrates the system's workflow in an industrial setting.

Sensors embedded in machines monitor critical parameters like vibration, temperature, speed, torque, and tool wear during operation. These sensors transmit real-time data to the cloud via secure, high-speed connections. The cloud provides the necessary computational resources to process large volumes of data. In the cloud, data undergoes preprocessing to clean and organize it, ensuring it is accurate and ready for analysis. Machine learning models then analyze this data to detect patterns, trends, and anomalies. Trained on historical data, these models predict potential faults by identifying early warning signs, such as abnormal temperature increases or irregular torque patterns. When a potential issue is detected, the system generates alerts for maintenance teams through dashboards, notifications, or emails. Alerts specify the affected machine or component, the type of fault, and its urgency, enabling technicians to address problems proactively. By preventing unexpected breakdowns, this system reduces downtime, cuts repair costs, and extends machine life. As more data is collected, the system improves its accuracy, making predictive maintenance an efficient and reliable alternative to traditional maintenance practices.

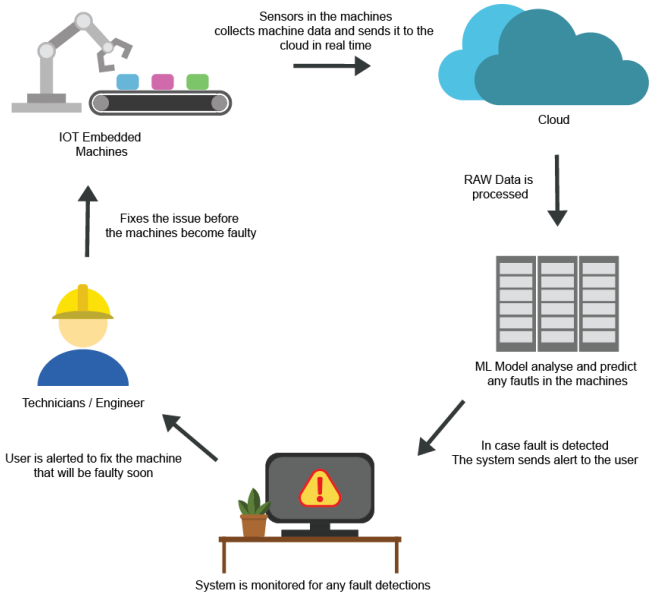


Fig. 1. System overview.

C. Methodology

This study follows the CRISP-DM methodology, a structured and iterative framework ideal for developing predictive maintenance systems [18]. Fig. 2 illustrates the entire methodology overview. It begins with data collection, using existing datasets to ensure high-quality and representative data. In data preprocessing, key steps include data cleansing, addressing imbalances, data smoothing, standardization, and splitting data into training (80%) and testing (20%) sets.

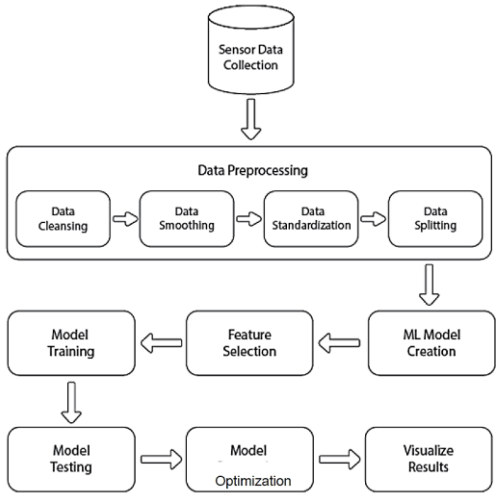


Fig. 2. Methodology overview.

During the model creation phase, the study replicates a prior model and develops three additional benchmark models. Feature importance is analyzed before training begins. The model testing phase evaluates performance on the test set using metrics like accuracy, precision, and F1 score to assess generalization capability. The proposed E.XGB model is then optimized through techniques like hyperparameter tuning and cross-validation to enhance accuracy and efficiency. Finally, visualizing results with tools such as confusion matrices and ROC curves provides insights into model performance and supports transparent communication with stakeholders.

D. Technical Setup and Development Environment

The simulation for this study was developed in Visual Studio Code (VS Code), a popular, free, and open-source code editor known for its versatility and extensive support for plugins [20]. The programming language used is Python, widely recognized for its effectiveness in data science and machine learning due to its powerful libraries [21].

Several Python libraries were employed for simulation. Pandas was used for data cleaning, preparation, and transformation [22]. Scikit-learn facilitated data pre-processing, machine learning model imports, feature selection, and performance metrics calculations. The Time library helped track training and processing time, while Matplotlib and Seaborn were used for visualizations.

The simulations were run on a system with the following specifications:

- Processor: Intel(R) Core (TM) i7-8750H CPU
- Memory: 32 GB RAM
- Operating System: Microsoft Windows 11 Home Single Language.

E. Dataset

Datasets are a structured collection of data used for analysis, modelling, and decision-making in various fields such as machine learning, statistics, and research. One of the key limitations of the previous approach is its reliance on a single dataset for experimentation [1, 2, 10, 11, 12, 13, 14, 15, 16, 19]. This restricts the generalizability of the model, as its performance has not been validated across diverse IoT environments with varying sensor data and operational conditions. In the context of IoT-based predictive maintenance, different datasets may exhibit variations in data distribution, sensor noise, and failure patterns, which a single dataset cannot fully capture. Without evaluation on multiple datasets, the robustness of the method remains uncertain, increasing the risk of overfitting to dataset-specific characteristics. Therefore, in our experiment, we use two datasets to enhance the generalizability and robustness of our proposed approach, ensuring its applicability across different IoT environments. The datasets used are the AI4I 2020 Dataset [8] and the CNC Milling Dataset [9], with each representing synthetic and real-world data, respectively. Additionally, these datasets exhibit both imbalanced and balanced characteristics, allowing us to assess the model's ability to handle class imbalance issues, which are common in real-world predictive maintenance applications.

1) AI4I2020 dataset: The AI4I 2020 Predictive Maintenance Dataset [8] is a synthetic dataset designed to replicate real-world industrial predictive maintenance data. It is widely used as a benchmark for machine learning algorithms, particularly for handling imbalanced datasets in predictive maintenance, quality assurance, and anomaly detection tasks. The dataset combines time-series and static data, including sensor readings, control signals, and process parameters, making it a valuable resource for developing predictive maintenance methods. The dataset contains 10,000 rows with two IDs, which are UID and Product ID, six features, and six failure modes. The features include product type, air temperature [K], process temperature [K], rotational speed [rpm], torque [Nm], and tool wear [min]. The six failure modes are machine failure, tool wear failure (TWF), heat dissipation failure (HDF), power failure (PWF), overstrain failure (OSF), and random failures (RNF).

2) CNC milling dataset: The CNC Milling Dataset from the University of Michigan SMART Lab [9] is a real-world dataset based on machining experiments conducted on a CNC milling machine. Data was collected from four machine motors, X, Y, Z axes, and spindle at a sampling rate of 100ms during 18 experiments on wax blocks. The dataset includes 25,286 rows with 55 features, including 1 ID and 3 failure modes as tool_condition, machining_finalized, and passed_visual_inspection. This study will use data from all 18 experiments but focus on selected features to challenge the model with limited input data. The tool_condition feature will be the target variable. Unlike the synthetic and imbalanced AI4I 2020 dataset, this balanced dataset provides an opportunity to evaluate the model's performance in a real-world environment, enhancing the robustness of the experiment.

IV. Enhanced Predictive Maintenance Method Using XGB

The proposed enhanced predictive maintenance method that is used in this study is designed based on the [1] approach to predictive maintenance for IOT industrial applications. However, the core model is modified along with several other minor changes in this study to further enhance the method proposed by [1]. This study's proposed enhanced predictive maintenance method aims to improve by eliminating some of the challenges present in that method.

Fig. 3 illustrates the flowchart of the entire enhanced predictive maintenance method. The flowchart provides a detailed pipeline for predictive fault detection using sensor data and an E.XGB model. This end-to-end process is designed to systematically prepare the data, refine feature selection, optimize the model, and deliver accurate predictions for identifying faults. By integrating feature importance methods, data preprocessing, and hyperparameter optimization, the pipeline aims to achieve reliable and efficient results, particularly for predictive maintenance applications.

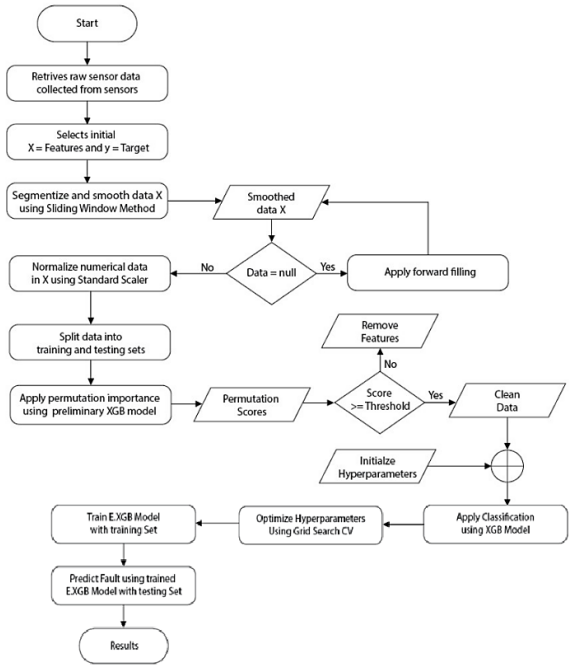


Fig. 3. Enhanced predictive method overview.

A. Data Pre-Processing

Pre-processing is the first and critical step in the machine learning workflow to ensure that the data is clean, consistent, and suitable for analysis. Raw data is often messy, incomplete, or unstructured, which can lead to inaccurate or inefficient model training. Preprocessing helps address these issues by transforming the data into a format that models can understand and learn effectively. After receiving the data collected from the sensors and the initial features x and target y selection is done, several steps are gone through to pre-process the data. Detailed information on how each pre-processing step is done is described in the sections below.

B. Sliding Window Method

One of the key enhancements in our proposed approach is the introduction of a data smoothing process using the Sliding Window Method. This technique is particularly well-suited for preprocessing IoT data, including streaming data, as demonstrated in [23]. By applying the Sliding Window Method to raw sensor data, we effectively smooth fluctuations and manage missing values, improving data consistency. The method segments the data into fixed-sized windows and computes the mean within each window, ensuring a more stable and reliable representation of the underlying patterns in the dataset.

The mean within each window starting at index t is computed as:

$$\mu_t = \frac{1}{K} \sum_{i=t}^{t+K-1} x_i$$

where,

μt is the mean of the window starting at position t.

Xi represents the data points in the window.

k is the number of elements in each window.

The summation runs from i=t to t+k−1

The missing values in the dataset are fixed using the forward filling method.

TABLE I. SLIDING WINDOW PARAMETERS

Parameter	Method	Inplace	Rolling	Window	Minimum required data points
Range	Forward filling	true	mean	3	1

Table I summarizes the parameters used for the Sliding Window Method in the proposed E.XGB approach. The forward fill (ffill) method replaces missing values with the last valid entry, ensuring data continuity. With *inplace* set to true, changes are made directly to the dataset. A rolling window of size 3 computes the mean over three consecutive values, moving forward one step at a time. The minimum required data points parameter is set to 1, ensuring that the mean is calculated even if some values are missing, as long as at least one valid value exists. This method smooths fluctuations while effectively handling missing data.

C. Data Normalization

The next step for data preprocessing in the proposed method is data scaling. During the data scaling stage, numerical data in the dataset are selected to undergo normalization. For this, the StandardScaler method is imported from Scikit-Learn's preprocessing library. The StandardScaler uses the z-score normalization technique to normalize the data. The z-score normalization technique is applied to each data point Di in category j, which calculates the number of standard deviations each point is away from the mean of its category. The z-score normalization is defined as:

$$D_j^{i'} = \frac{D_j^i - \mu_j}{\sigma_j}$$

where,

$D_j^{i'}$ is the normalized data point for feature i in category j,

D_j^i is the original data point in category j,

μ_j is the mean of data category j,

σ_j is the standard deviation of data category j.

Normalization ensures that all features are on a similar scale, centered at the origin, with a peak of 1. This improves the stability of training by reducing variance and the impact of outliers. This method ensures that features have a mean of 0 and a standard deviation of 1, making the model more robust to outliers.

D. Data Splitting

After normalization, the dataset is split into training and testing sets to ensure the model trains on one subset and is tested on unseen data, reducing bias. Without splitting, testing on the same dataset used for training could lead to misleading results.

TABLE II. DATA SPLIT PARAMETERS

Parameter	Train size	Test size	Random state
Value	0.8	0.2	42

The train_test_split function from Scikit-Learn's model selection library is used for this purpose, with parameters specified in Table II. For this study, 80% of the data is allocated for training and 20% for testing. The data is shuffled randomly before splitting, with a random state of 42 set to ensure reproducibility.

E. Feature Selection

After splitting the data, feature selection is performed using permutation importance, a technique that evaluates the significance of each feature by shuffling its values and observing the impact on model performance, with a significant drop in accuracy indicating importance. The permutation importance function from Scikit-Learn's inspection library is applied to the preliminary XGB model and dataset to calculate a reference score. The importance i_j for feature f_j is calculated as follows:

$$i_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j}$$

where,

K is the total number of data points in the dataset,

k is the k^{th} datapoint.

For each feature, the values are shuffled to create a distorted dataset, and the model is re-scored to determine feature importance based on the performance drop. The process uses the number of repeats parameter to specify the number of shuffles, where higher values provide more reliable scores but increase computation time.

TABLE III. PERMUTATION IMPORTANCE PARAMETERS

Parameter	Number of estimators	Maximum depth	Learning rate	Number of repeats
Value	100	10	0.05	10

Table III shows the importance of the permutation parameters used. The initial XGB model is configured with key hyperparameters of number of estimators to 100 trees, maximum depth to 10 to balance complexity and overfitting, and learning rate of 0.05, to control tree contribution. Features with low importance scores, based on a predefined threshold, are removed to streamline the dataset, retaining only the most relevant features and improving model efficiency and performance.

F. XGB Classification

Once data preprocessing is complete, the core XGB model is initialized using the XGBoost classifier from the *xgboost* library. XGB is a highly efficient implementation of the gradient boosting framework, known for its speed, accuracy, and flexibility in handling structured data. It builds an ensemble of decision trees sequentially, optimizing each tree to correct errors from the previous ones using gradient descent to minimize the loss function. The training process starts with an initial prediction, often the mean of the target values. The objective function combines a loss function to measure prediction errors and a regularization term to prevent overfitting. The function is:

$$Obj(\theta) = \sum_{i=1}^{m} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where,

$L\ (y_i, \hat{y}_i)$ is the loss function, which is the difference between true and predicted values,

$\Omega\ (f_k)$ is the regularization term for the weak learner f_k,

m is the number of data points.

K is the number of trees.

At each iteration, the model adds a new tree, updating the prediction. The final prediction is the sum of all tree predictions. XGB excels at handling large datasets and complex relationships efficiently. Its combination of speed, accuracy, and regularization makes it a top choice for structured data and a frequent winner in machine learning competitions.

G. Hyperparameter Optimization Using GridSearchCV

GridSearchCV is used to fine-tune the hyperparameters of the XGB model by testing all possible combinations of predefined values to find the best set. It evaluates each combination using 3-fold cross-validation, ensuring the model performs well on unseen data while avoiding overfitting. This process improves both accuracy and robustness. Table IV shows the ranges for key hyperparameters set and Table V shows the GridSearchCV parameters used.

TABLE IV. XGB HYPERPARAMETER SEARCH RANGE

Parameter	Number of estimators	Maximum depth	Learning rate	Sub sample
Range	50 - 200	1 - 10	0.01 - 0.07	0.01 - 0.07

TABLE V. GRIDSEARCHCV PARAMETERS

Parameter	Model	Scoring	CV
Value	XGBClassifier (eval_metric='logloss')	accuracy	3

For this study, GridSearchCV is imported from Scikit-Learn's model selection library and applied to the XGB classifier, initialized with an evaluation metric set based on the Logarithmic Loss value. The search grid includes ranges for key hyperparameters such as the number of estimators at 50 to 200, maximum depth at 1 to 10, learning rate at 0.01 to 0.07, and subsample at 0.01 to 0.07. These parameters influence the model's performance, and GridSearchCV compares average accuracy scores across all combinations to select the best one. The optimal hyperparameters are then used to build the final XGB model.

V. RESULTS AND DISCUSSION

This section summarizes the experimental results on the AI4I2020 dataset and CNC Milling dataset, comparing various

machine learning models across key metrics such as accuracy, precision, recall, F1-score, ROC-AUC, confusion matrices, training time, and prediction time. The results are organized into tables and visualized with figures to provide a comprehensive view of model performance and to facilitate comparisons.

A. Pre-processing on AI4I2020 Dataset

The AI4I2020 dataset underwent a series of preprocessing steps to ensure quality and consistency. Initially, the dataset was checked for missing values, and no null entries were found. Irrelevant features such as UDI, Product ID, and Type were removed, as they did not contribute to machine failure predictions, leaving air temperature, process temperature, rotational speed, torque, and tool wear as the selected features. The study also focused on predicting overall machine failures rather than specific failure types, so failure mode features like TWF, HDF, and PWF were excluded.

Fig. 4. Data balance in AI4I2020 dataset.

As shown in Fig. 4, the dataset revealed a significant class imbalance, with only 3.4% of instances labelled as machine failures. The data was then smoothed and segmented using a sliding window method. To address the varying scales of features, standardization was applied, ensuring all features had a mean of 0 and a standard deviation of 1, making them comparable and enhancing model performance. Finally, the data was split into 80% training and 20% testing subsets, preparing it for machine learning tasks.

TABLE VI. PARAMETER OPTIMIZATION RESULT ON AI4I2020 DATASET

Parameter	Number of estimators	Maximum depth	Learning rate	Sub sample
Optimum	200	8	0.07	0.5

Hyperparameter optimization for the XGB model was performed using GridSearchCV, resulting in an enhanced model configuration. The optimal parameters are shown in Table VI, which includes the number of estimators set to 200, maximum depth set to 8, learning rate set to 0.07, and subsample set to 0.5. These parameters were carefully chosen to balance model complexity and generalization, ensuring robust performance. A higher maximum depth allowed the model to capture more complexity without overfitting, while a smaller learning rate enabled gradual and stable learning. The subsample value helped mitigate overfitting by training the model on random subsets of data at each iteration. With these optimized hyperparameters, the E.XGB model achieved significant improvements in metrics such as accuracy, precision, recall, and F1-score, demonstrating its suitability for predicting machine failures in the AI4I2020 dataset.

B. Model Results on AI4I2020 Dataset

The classification performance of various models on the AI4I2020 dataset is summarized in Fig. 5 and Table VII, comparing E.XGB, E.AB, RF, DT, and SVM based on accuracy, precision, recall, F1-score, ROC AUC, training time, and prediction time. Among all models, the proposed E.XGB model outperforms the others, achieving the highest accuracy of 99.05%, precision of 93.62%, F1-score of 82.25%, and ROC AUC of 97.10%.

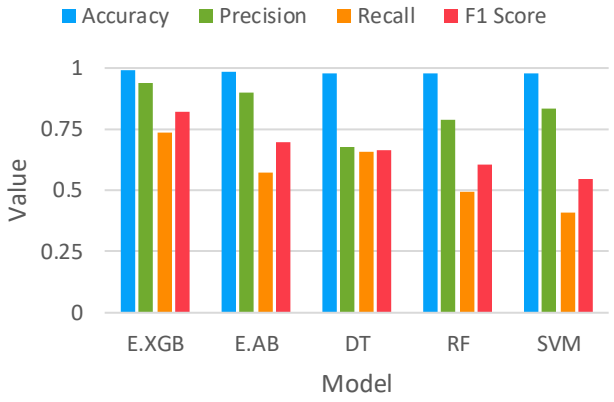


Fig. 5. Efficiency results on CNC AI4I2020 dataset.

Compared to E.AB, RF, DT, and SVM, the E.XGB demonstrates superior predictive capability while maintaining efficiency. Although DT achieves a recall of 65.57%, slightly lower than E.XGB's 73.33%, it lags in precision at only 67.80%, making it less reliable overall.

TABLE VII. PERFORMANCE METRICS OF MODELS ON AI4I2020 DATASET

Details	E.XGB	E.AB	DT	RF	SVM
Accuracy (%)	99.05	98.50	98.00	98.05	97.95
Precision (%)	93.62	89.74	67.80	78.95	83.33
Recall (%)	73.33	57.38	65.57	49.18	40.98
F1-Score (%)	82.25	70.00	66.67	60.61	54.95
ROC AUC (%)	97.10	94.64	87.26	86.63	85.35
Training Time (s)	0.2102	3.9122	0.0312	0.4531	0.4343
Prediction Time (s)	0.0001	0.0544	0.0001	0.0189	0.0937

In terms of computational performance, E.XGB has a training time of 0.2102s, significantly faster than E.AB at 3.9122s and RF at 0.4531s, while being slightly slower than DT at 0.0312s. For prediction time, E.XGB and DT are the fastest at 0.0001s, making them highly efficient for real-time applications. Although DT is computationally fast, it suffers from lower precision and an overall weaker predictive performance. Overall, for the AI4I2020 dataset, E.XGB proves to be the most effective model, offering the best trade-off between accuracy, precision, recall, and efficiency. DT, despite

being computationally efficient, falls short in terms of precision, making E.XGB is the most suitable choice for predictive maintenance applications in IoT-embedded machinery.

C. Pre-processing on CNC Milling Dataset

The CNC Milling Dataset was pre-processed to ensure high-quality data for analysis. Missing values were checked and found to be absent, confirming data completeness. Feature selection was then performed, and a set of relevant features was chosen based on preliminary importance tests. The selected features included x1_actualposition, y1_actualposition, z1_actualposition, s1_actualposition, and m1_current_feedrate, with tool condition as the target feature, which directly relates to machinery degradation.

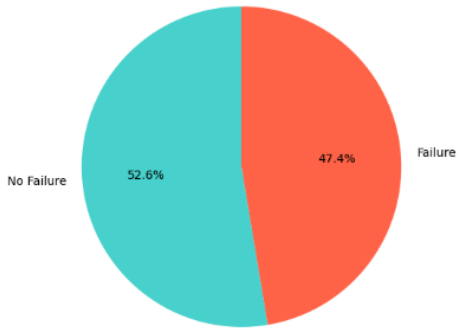


Fig. 6. Data balance in CNC milling dataset.

Unlike the AI4I2020 dataset, the CNC Milling dataset exhibited a balanced class distribution, as shown in Fig. 6, with 47.4% labelled as machine failures. The data is then smoothed and segmented using the sliding window method. The dataset also revealed significant differences in feature scales, with features like X1_ActualPosition spanning wide ranges. To address this, all features were standardized, ensuring equal contribution to model performance. The final step was splitting the data into 80% training and 20% testing sets, preparing it for model training and evaluation.

TABLE VIII. PARAMETER OPTIMIZATION RESULT ON CNC MILLING DATASET

Parameter	Number of estimators	Maximum depth	Learning rate	Sub sample
Optimum	200	10	0.07	0.7

GridSearchCV was applied to the base XGB model to find the optimal hyperparameters. The best combination of parameters found is as shown in Table VIII, where the number of estimators is set to 200, the maximum depth set to 10, the learning rate set to 0.07, and the subsample set to 0.7. These parameters were chosen to strike a balance between model complexity and generalization. The maximum depth of 10 allowed the model to capture sufficient complexity, while the learning rate of 0.07 ensured gradual learning. The subsample value of 0.7 helped reduce overfitting by training the model on a randomly selected subset of the data. These hyperparameters improved the model's performance, as evidenced by better evaluation metrics such as accuracy, precision, recall, and F1-score.

D. Model Results on CNC Milling Dataset

The classification performance of various models on the CNC Milling dataset is summarized in Fig. 7 and Table IX, comparing E.XGB, E.AB, RF, DT, and SVM based on accuracy, precision, recall, F1-score, ROC AUC, training time, and prediction time. Among these models, the proposed E.XGB demonstrates superior performance, achieving the highest accuracy of 99.01%, precision of 99.57%, recall of 98.30%, F1-score of 98.93%, and ROC AUC of 99.98%.

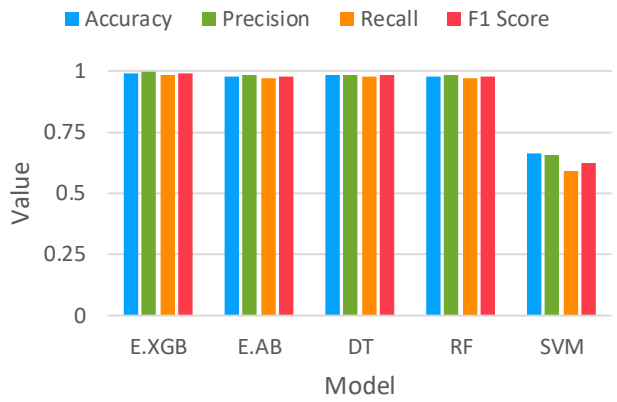


Fig. 7. Efficiency results on CNC milling dataset.

Compared to E.AB, RF, DT, and SVM, E.XGB consistently delivers higher predictive capability while maintaining computational efficiency. In terms of training time, E.XGB completes training in just 0.3126s, significantly outperforming E.AB at 4.6340s, RF at 1.8848s, and SVM at 81.1611s, while being slightly slower than DT at 0.0473s. Similarly, for prediction time, E.XGB and DT are the fastest at 0.0001s, far surpassing the efficiency of other models, particularly SVM, which has the slowest prediction time at 6.8742s.

TABLE IX. PERFORMANCE METRICS OF MODELS ON CNC MILLING DATASET

Details	E.XGB	E.AB	DT	RF	SVM
Accuracy (%)	99.01	97.71	98.16	97.88	66.53
Precision (%)	99.57	98.19	98.38	98.62	65.72
Recall (%)	98.30	96.86	97.67	96.82	58.95
F1-Score (%)	98.93	97.52	98.02	97.71	62.15
ROC AUC (%)	99.98	99.71	99.64	99.87	72.29
Training Time (s)	0.3126	4.6340	0.0473	1.8848	81.1611
Prediction Time (s)	0.0001	0.0696	0.0001	0.0781	6.8742

Despite DT being the fastest to train and predict, it still falls short in terms of accuracy of 98.16%, precision of 98.38%, and recall of 97.67%, making it less reliable for high-accuracy predictive maintenance applications. In contrast, SVM struggles significantly with an accuracy of only 66.53% and the lowest recall of 58.95%, proving unsuitable for this dataset.

Overall, E.XGB once again stands out as the most effective model for predictive maintenance in IoT-embedded machinery, offering the best balance of accuracy, precision, recall, and efficiency.

VI. Conclusion

In conclusion, this study proposes an enhanced predictive maintenance method for IoT-embedded industrial machinery, addressing challenges related to prediction accuracy, training efficiency, and prediction latency. The core of the method is the E.XGB model, which integrates enhanced feature selection, preprocessing, and hyperparameter optimization to overcome limitations of traditional approaches. The model demonstrated superior performance on the AI4I2020 and CNC Milling datasets, excelling in accuracy, precision, recall, and F1-scores, crucial for minimizing false positives and negatives in fault detection. It achieved 99.05% accuracy on the AI4I 2020 dataset and 99.01% accuracy on the CNC Milling dataset. Moreover, the proposed model demonstrated training speed of approximately 94% and prediction speeds of approximately 99.8% improvement than the E.AB model, thus meeting real-time industrial application requirements.

The optimized training and prediction pipelines resulted in significant reductions in computational time, making the model suitable for real-time applications. A comparative analysis with the E.AB method from previous research, as well as baseline models DT, RF, and SVM, highlighted the E.XGB model's consistent superiority. Its combination of advanced optimization techniques and lightweight design enables scalability across diverse operational conditions. The primary goal of the research which is enhancing predictive maintenance was achieved by improving accuracy, training efficiency, and prediction speed. Permutation importance and hyperparameter tuning enabled the identification of critical features for precise fault detection. Optimized preprocessing and grid search tuning reduced training time, while the lightweight architecture ensured low-latency predictions essential for real-time decision-making.

The proposed method minimizes unplanned downtime, reduces maintenance costs, and extends machinery lifespan, with scalability and adaptability for various industries. By enhancing predictive accuracy, training speed, and prediction latency, this study bridges the gap between theoretical machine learning advancements and their practical applications in industrial maintenance, contributing to a more efficient and sustainable industrial ecosystem.

Future research could focus on extending the proposed E.XGB model towards integration with IoT and edge devices, enabling deployment in resource-constrained environments for real-time predictive maintenance. In addition, exploring hybrid approaches that combine E.XGB with deep learning techniques such as CNNs or LSTMs may further enhance the model's ability to capture complex temporal and nonlinear fault patterns. To ensure adoption in real industrial settings, incorporating explainable AI (XAI) methods will be essential for improving model interpretability and building trust among engineers and decision-makers.

References

[1] Elkateb, S., Métwalli, A., Shendy, A., & Abu-Elanien, A. E. (2024). Machine learning and IoT – Based predictive maintenance approach for industrial applications. Alexandria Engineering Journal /Alexandria Engineering Journal, 88, 298–309. https://doi.org/10.1016/j.aej.2023.12.065

[2] Lee, W. J., Wu, H., Yun, H., Kim, H., Jun, M. B., & Sutherland, J. W. (2019). Predictive maintenance of machine tool systems using artificial intelligence techniques applied to machine condition data. Procedia CIRP, 80, 506–511. https://doi.org/10.1016/j.procir.2018.12.019

[3] Wang, L., Zhu, Z., & Zhao, X. (2024). Dynamic predictive maintenance strategy for system remaining useful life prediction via deep learning ensemble method. Reliability Engineering & Systems Safety, 110012. https://doi.org/10.1016/j.ress.2024.110012

[4] Meriem, H., Nora, H., & Samir, O. (2023). Predictive Maintenance for Smart Industrial Systems: a Roadmap. Procedia Computer Science, 220, 645–650. https://doi.org/10.1016/j.procs.2023.03.082

[5] Dalzochio, J., Kunst, R., Pignaton, E., Binotto, A., Sanyal, S., Favilla, J., & Barbosa, J. (2020). Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges. Computers in Industry, 123, 103298. https://doi.org/10.1016/j.compind.2020.103298

[6] Hamaide, V., Joassin, D., Castin, L., & Glineur, F. (2022). A two-level machine learning framework for predictive maintenance: Comparison of learning formulations. arXiv preprint arXiv:2204.10083. Retrieved from https://arxiv.org/abs/2204.10083

[7] Chevtchenko, S. F., dos Santos, M. C. M., Vieira, D. M., Mota, R. L., Rocha, E., Cruz, B. V., Araújo, D., & Andrade, E. (2023). Predictive maintenance model based on anomaly detection in induction motors: A machine learning approach using real-time IoT data. arXiv preprint arXiv:2310.14949. Retrieved from https://arxiv.org/abs/2310.14949

[8] Matzka, S. (2020). UCI machine learning repository, AI4I 2020 predictive maintenance

[9] Sun, S. (2020). University of Michigan System-level Manufacturing and Automation Research Testbed (SMART) Lab, CNC Mill Tool Wear Detection Dataset.

[10] Arisani, M. I., & Muljono, M. (2024). Bagging Nearest Neighbor and its Enhancement for Machinery Predictive Maintenance. Journal of Applied Informatics and Computing, 8(2), 248–256. https://doi.org/10.30871/jaic.v8i2.8158

[11] Ghadekar, P., Manakshe, A., Madhikar, S., Patil, S., Mukadam, M., & Gambhir, T. (2024). Predictive Maintenance for Industrial Equipment: Using XGBoost and Local Outlier Factor with Explainable AI for analysis (pp. 25–30). https://doi.org/10.1109/confluence60223.2024.10463280

[12] Baldovino, R. G., Camacho, K. S. I., Chua-Unsu, M. V. H. Y., Go, J. L. C., Munsayac, F. E. T., & Bugtai, N. T. (2024). Comparative Analysis of Machine Learning Models for Predictive Analysis of Machine Failures. In 2024 9th International Conference on Mechatronics Engineering (ICOM) (Vol. 10, pp. 288–293). https://doi.org/10.1109/icom61675.2024.10652325

[13] Elminir, H. K., El-Brawany, M. A., Ibrahim, D. A., Elattar, H. M., & Ramadan, E. (2024). An efficient deep learning prognostic model for remaining useful life estimation of high-speed CNC milling machine cutters. Results in Engineering, 24, 103420. https://doi.org/10.1016/j.rineng.2024.103420

[14] Kaliyannan, D., Thangamuthu, M., Pradeep, P., Gnansekaran, S., Rakkiyannan, J., & Pramanik, A. (2024). Tool condition monitoring in the milling process using deep learning and reinforcement learning. Journal of Sensor and Actuator Networks, 13(4), 42. https://doi.org/10.3390/jsan13040042

[15] Kamat, P. V., Kumar, S., Patil, S., Sugandhi, R., & Nargund, A. (2022). Tool Wear Prediction in Milling: A comparative analysis based on machine learning and deep learning approaches. International Journal of Computing and Digital Systems, 11(1), 151–165. https://doi.org/10.12785/ijcds/110112

[16] Assagaf, I., Sukandi, A., Abdillah, A. A., Arifin, S., & Ga, J. L. (2023). Machine predictive maintenance by using support vector machines. Recent in Engineering Science and Technology, 1(01), 31–35. https://doi.org/10.59511/riestech.v1i01.6

[17] Misaii, H., Fouladirad, M., Askari, B., & Durupt, A. (2024). Predictive Degradation Modelling Using Artificial Intelligence: Milling Machine Case Study. Advances in Reliability, Safety and Security.

[18] Stow, M. T. (2024). Hybrid Deep learning approach for predictive maintenance of industrial machinery using convolutional LSTM networks. International Journal of Computer Sciences and Engineering, 12(4), 1–11. https://doi.org/10.26438/ijcse/v12i4.111

[19] Gong, L., & Chen, Y. (2024). Machine Learning-enhanced loT and Wireless Sensor Networks for predictive analysis and maintenance in wind turbine systems. International Journal of Intelligent Networks, 5, 133–144. https://doi.org/10.1016/j.ijin.2024.02.002

[20] Richardson, B. (2019). Visual Studio Code as a Lightweight IDE: Features and Flexibility. Journal of Open Source Software Tools and IDEs, 3(1), 87-91.

[21] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). TensorFlow: A system for large-scale machine learning. 12th {USENIX} Symposium on Operating Systems Design and Implementation (OSDI} 16), 265-283.

[22] McKinney, W. (2010). Data structures for statistical computing in Python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51-56).

[23] Raja Azhan Syah Raja Wahab, Siti Nurulain Mohd Rum, Hamidah Ibrahim, Iskandar Ishak, "Efficient Top-K Continuous Query Processing Over Sliding Window Model (SWM) Method on Uncertain Data Stream," WSEAS Transactions on Systems and Control, vol. 19, pp. 283-308, 2024, DOI:10.37394/23203.2024.19.31.

A Hybrid Deep Learning and Forensic Approach for Robust Deepfake Detection

Sales Aribe Jr.
Information Technology Department, Bukidnon State University, Malaybalay City, Philippines

***Abstract*—The rapid evolution of generative adversarial networks (GANs) and diffusion models has made synthetic media increasingly realistic, raising societal concerns around misinformation, identity fraud, and digital trust. Existing deepfake detection methods either rely on deep learning, which suffers from poor generalization and vulnerability to distortions, or forensic analysis, which is interpretable but limited against new manipulation techniques. This study proposes a hybrid framework that fuses forensic features—including noise residuals, JPEG compression traces, and frequency-domain descriptors—with deep learning representations from convolutional neural networks (CNNs) and vision transformers (ViTs). Evaluated on benchmark datasets (FaceForensics++, Celeb-DF v2, DFDC), the proposed model consistently outperformed single-method baselines and demonstrated superior performance compared to existing state-of-the-art hybrid approaches, achieving F1-scores of 0.96, 0.82, and 0.77, respectively. Robustness tests demonstrated stable performance under compression (F1 = 0.87 at QF = 50), adversarial perturbations (AUC = 0.84), and unseen manipulations (F1 = 0.79). Importantly, explainability analysis showed that Grad-CAM and forensic heatmaps overlapped with ground-truth manipulated regions in 82 per cent of cases, enhancing transparency and user trust. These findings confirm that hybrid approaches provide a balanced solution—combining the adaptability of deep models with the interpretability of forensic cues—to develop resilient and trustworthy deepfake detection systems.**

***Keywords*—Adversarial robustness; deepfake detection; diffusion models; explainable AI; forensic fusion; multimedia forensics; trustworthy AI**

I. Introduction

The emergence of generative artificial intelligence (AI) models, such as GANs and diffusion models, has dramatically advanced the creation of synthetic media. These technologies, popularly associated with deepfakes, produce hyper-realistic images and videos that are increasingly difficult to distinguish from authentic content. Deepfake content is proliferating rapidly, with recent reports estimating that deepfake files surged from around 500,000 in 2023 to over 8 million in 2025, business losses an average of nearly $500,000 per deepfake-related incident, and human detection rates on high-quality deepfake videos dropping to 24.5 per cent, especially when compression or post-processing is involved [1]. While deepfakes have legitimate applications in film production, education, and digital creativity, they also pose serious societal risks, including misinformation, identity fraud, political manipulation, and erosion of public trust in digital media [2], [3].

Current deepfake detection research has largely focused on deep learning approaches, particularly CNNs, recurrent neural networks (RNNs), and, more recently, transformer-based architectures. These models learn discriminative representations that can separate authentic from manipulated content. Despite their success, they suffer from several limitations: poor generalization across datasets, sensitivity to compression or adversarial attacks, and a lack of transparency in decision-making [4], [5]. In contrast, classical forensic analysis—based on principles of image formation and physical signal processing—examines sensor noise patterns, JPEG compression artifacts, or frequency inconsistencies. These techniques are interpretable and reliable for certain manipulations, but they often fail to adapt to new generation techniques [6], [7].

The shortcomings of purely deep learning or purely forensic approaches have spurred interest in hybrid frameworks that integrate both [8]. Hybrid methods combine the adaptability of learned representations with the interpretability of forensic traces, yielding systems that are more robust and explainable [9], [10]. For instance, ensembles of CNNs and transformers capture both local pixel-level cues and global temporal context, while forensic features provide interpretable indicators such as abnormal noise residuals or disrupted frequency patterns. Recent studies suggest that such fusion can mitigate the challenges of cross-dataset generalization and robustness to real-world distortions [11], [12], [13], [14].

Nevertheless, several gaps remain. First, many detection systems continue to overfit to specific datasets, limiting their ability to generalize to novel manipulation techniques. Second, most detectors exhibit performance degradation under realistic video conditions such as compression, resolution changes, and adversarial noise. Third, explainability remains underexplored; stakeholders such as law enforcement agencies, media organizations, and social platforms increasingly require interpretable justifications for automated decisions. Finally, resource efficiency—critical for real-time deployment—remains an open challenge.

This study addresses these challenges by proposing a hybrid deepfake detection framework that fuses forensic features with deep learning representations. Specifically, the approach integrates noise residuals, compression inconsistencies, and frequency-domain descriptors with CNN and transformer-based features through a fusion architecture. The model is evaluated across multiple datasets, manipulation types, and compression levels, with emphasis on cross-dataset generalization and robustness under adversarial settings. In addition, interpretability is enhanced by visualizing manipulated regions

and forensic traces, supporting explainable AI in multimedia forensics.

The key contributions of this study are fourfold. First, it proposes a hybrid deepfake detection framework that uniquely integrates forensic cues—including noise residuals, JPEG compression traces, and frequency-domain descriptors—with deep learning representations from convolutional neural networks and vision transformers through a dedicated fusion architecture. Second, the proposed model emphasizes robustness and cross-dataset generalization, addressing the persistent limitations of existing detectors that struggle with unseen manipulations, compression, and adversarial distortions. Third, the study incorporates an explainability module that fuses Grad-CAM visualizations with forensic heatmaps to enhance the transparency and interpretability of detection results. Finally, this research bridges the gap between AI adaptability and forensic interpretability, contributing to the development of more trustworthy, resilient, and explainable deepfake detection systems.

II. Related Work

A. Deep Learning-Based Approaches

Most state-of-the-art deepfake detectors rely on deep neural networks to automatically learn discriminative features from data. CNNs have been widely adopted due to their ability to capture spatial artifacts in manipulated images and videos. For instance, Afchar et al. [15] proposed MesoNet, a compact CNN architecture designed for video forgery detection, demonstrating strong performance on low-resolution inputs. Similarly, Rossler et al. [16] introduced the FaceForensics++ benchmark and showed that CNN-based models trained on large-scale datasets can detect multiple manipulation techniques.

Beyond CNNs, RNNs, and Long Short-Term Memory (LSTM) models have been applied to capture temporal inconsistencies in videos [17]. More recently, transformer-based architectures have gained traction due to their ability to model long-range dependencies. ViT-based models and spatio-temporal transformers have been shown to outperform CNNs in cross-dataset evaluations [18]. Despite their success, purely deep learning-based methods suffer from limited interpretability and often fail to generalize across unseen manipulation techniques, particularly when trained on specific datasets [5].

B. Forensic-Based Approaches

Unlike deep learning methods, forensic approaches are grounded in principles of digital image formation and signal processing. These methods exploit the fact that manipulations often disrupt underlying statistical or physical patterns of real images. For example, Photo Response Non-Uniformity (PRNU) has been used to detect inconsistencies in sensor noise, which can reveal tampered regions [19]. Similarly, Anwar et al. [20] showed that local descriptors derived from residual noise can be repurposed as CNN filters for forgery detection.

Compression-based forensic features, such as analyzing JPEG block artifacts and quantization inconsistencies, have also been widely explored [21]. Frequency-domain analysis, particularly through Discrete Cosine Transform (DCT) and wavelet features, provides another line of defense against synthetic content [22]. These forensic cues offer interpretability and robustness to small perturbations, but they often lack adaptability to new manipulation techniques, especially as generative models become more sophisticated.

C. Hybrid Approaches

To overcome the limitations of purely data-driven or purely forensic strategies, hybrid approaches have emerged that combine both. Mancy et al. [23] and Fardin et al. [18] proposed a hybrid model that integrates handcrafted forensic features with CNN-based deep features, showing improved robustness to compression and unseen manipulations. Chen et al. [25] developed a generalizable detector by fusing spatial forensic cues with learned representations, achieving competitive performance across multiple datasets.

Hybrid models leverage the strengths of both worlds. Deep learning ensures adaptability to evolving generative methods, while forensic features provide interpretability and resilience under distortions. For example, Haliassos et al. [11] introduced an audio-visual hybrid framework that detected mismatches between lip movements and speech signals, outperforming unimodal methods. Similarly, Dang et al. [26] demonstrated that fusing frequency-domain features with CNN embeddings significantly enhances cross-dataset generalization.

D. Research Gap

While progress has been made, several gaps remain. First, many detectors overfit to specific datasets and manipulation techniques, limiting real-world applicability. Second, robustness under practical conditions such as compression, resolution degradation, and adversarial attacks remains underexplored. Third, explainability in hybrid frameworks is still underdeveloped, despite increasing demand from forensic analysts, regulators, and end-users. These challenges motivate the proposed hybrid framework, which fuses deep features with forensic cues while explicitly emphasizing cross-dataset generalization, robustness, and interpretability.

III. Methodology

This study proposes a hybrid detection framework that integrates forensic analysis with deep learning models. The methodology is organized into seven components: dataset selection, preprocessing, feature extraction, hybrid model architecture, training protocol and validation, real-time deployment setup, and evaluation strategy.

A. Datasets

To achieve robustness and cross-dataset generalization, three benchmark datasets were employed. FaceForensics++ [16], a widely used benchmark which provides a large-scale collection of manipulated videos generated using various multiple manipulation techniques, while Celeb-DF v2 [27] offered high-quality celebrity deepfakes that are particularly challenging for detectors. The DeepFake Detection Challenge (DFDC) dataset [24] [5] contributed more than 100,000 clips exhibiting a wide range of real-world variations. Collectively, these datasets ensured both diversity and realism in evaluating the proposed approach. Thus, this combination enables the evaluation of both in-distribution and out-of-distribution performance, addressing generalization concerns.

B. Preprocessing

All video data were decomposed into individual frames at a fixed sampling rate valued at 25 fps. Frames were resized to 224 × 224 pixels and normalized to the [0,1] range. To preserve compression artifacts, which serve as important forensic cues, the frames were stored under artifact-preserving JPEG compression at varying quality sensor-level noise patterns of 50, 75, and 100. This design choice contrasts with standard preprocessing methods that typically remove compression traces, thus demonstrating the model's focus on maintaining forensic fidelity for improved manipulation detection. Face alignment was applied using the Multi-task Cascaded CNN (MTCNN), guaranteeing consistent cropping and alignment across frames to reduce variability unrelated to manipulation.

C. Feature Extraction

The framework draws upon both forensic and deep learning-based features. Forensic analysis was performed by extracting noise residuals through PRNU, which captures inconsistencies in sensor noise [28]; by analyzing JPEG compression patterns through block-level quantization artifacts to detect anomalies in compression traces [29]; and by deriving frequency-domain descriptors such as DCT coefficients and spectral distributions [22], which reveal spectral anomalies introduced during generation. These handcrafted features provided interpretable signals of manipulation.

In parallel, deep learning features were obtained using two complementary models. A ResNet-50 backbone [30] extracted local pixel-level features sensitive to subtle artifacts, while a Vision Transformer [31] captured long-range dependencies and temporal coherence across frames. The combination of CNN-based and transformer-based representations allowed the system to learn both fine-grained and holistic patterns of manipulation.

D. Hybrid Model Architecture

The proposed hybrid architecture fuses forensic and deep learning features through a dedicated fusion layer that enables joint learning, as shown in Fig. 1. Unlike conventional approaches that independently apply deep learning or forensic analysis, the proposed framework integrates both feature domains within a unified learning process. The fusion layer is not a simple concatenation but a trainable joint-representation space that learns correlations between forensic cues and deep model embeddings. This integration enables complementary feature learning that enhances robustness and interpretability, representing a methodological advancement over existing single-model detectors.

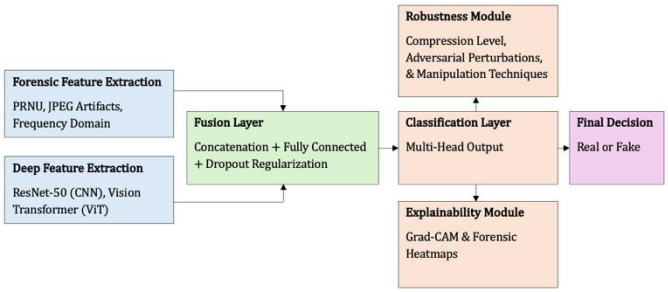


Fig. 1. Proposed hybrid deepfake detection model architecture.

In this framework, forensic and deep features are first extracted independently (feature extraction), after which they are concatenated and passed through a fully connected layer with dropout regularization to reduce overfitting (fusion layer). The fused representation is then processed by a multi-head classification block that outputs the probability of an input being real or manipulated (classification layer). To enhance interpretability, the architecture incorporates an explainability module that generates Gradient-weighted Class Activation Mapping (Grad-CAM) visualizations from deep features alongside forensic heatmaps derived from handcrafted cues (explainability layer), including a robustness test to examine model stability under practical distortions. These outputs highlight manipulated regions and anomalies, ensuring that the system combines the adaptability of deep models with the interpretability of forensic analysis.

E. Training Protocol and Validation

Models were trained using the Adam optimizer with an initial learning rate of 1×10^{-4}, decayed using a cosine annealing scheduler. Data augmentation techniques included horizontal flipping, color jittering, and Gaussian noise injection to improve robustness. Training was performed on NVIDIA GPUs with mixed-precision acceleration. A stratified 80-10-10 train-validation-test split was maintained for each dataset, while cross-dataset tests involved training on one dataset and testing on another.

F. Real-Time Deployment Setup

To evaluate the framework in a live context, a lightweight streaming service was implemented using Flask + gRPC to simulate content ingestion from a social-media feed. Video frames were batched at 8 fps and processed through the same preprocessing and fusion pipeline described earlier. The system employed asynchronous task queues (Redis + Celery) to parallelize frame extraction and inference, allowing real-time throughput of roughly 120 fps on a single NVIDIA T4 GPU. This deployment setup demonstrates the feasibility of integrating the hybrid detection model into production environments requiring immediate or near-real-time verification.

G. Evaluation Metrics

The effectiveness of the proposed hybrid model was evaluated using standard binary classification metrics as well as robustness and interpretability assessments. Let TP denote true positives, TN true negatives, FP false positives, and FN false negatives.

Accuracy measures the overall proportion of correctly classified samples:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Precision evaluates the reliability of positive predictions, i.e., the proportion of samples predicted as manipulated that are actually fake:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall (Sensitivity) assesses the detector's ability to identify manipulated samples:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-Score is the harmonic mean of precision and recall, balancing false positives and false negatives:

$$F1\text{-}Score = 2\ x\ \frac{Precision\ x\ Recall}{Precision + Recall} \quad (3)$$

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) evaluates discrimination capability across thresholds. It is defined as:

$$AUC = \int_0^1 TPR(FPR)\, d(FPR) \quad (5)$$

$$\text{where, } TPR = \frac{TP}{TP + FN} \text{ and } FPR = \frac{FP}{FP + TN}$$

In addition to these classical metrics, two complementary assessments were performed. Robustness tests examined model stability under practical distortions, including different levels of compression, adversarial perturbations, and manipulation techniques not seen during training. Explainability assessment provided qualitative insights by analyzing Grad-CAM visualizations and forensic heatmaps, highlighting manipulated facial regions and verifying that the model's decisions were grounded in interpretable evidence.

IV. Results and Discussion

A. Preprocessing Results

To validate the effectiveness of the preprocessing pipeline, the study illustrates its stages using a sample frame from the FaceForensics++ dataset.

Fig. 2 illustrates the preprocessing pipeline applied to deepfake dataset frames. Fig. 2(a) shows the original input frame. Fig. 2(b) depicts the aligned face obtained via face detection using MTCNN. Fig. 2(c) shows the standardized resized frame (224×224 pixels). Fig. 2(d) demonstrates artifact-preserving JPEG compression, which retains block-level distortions useful for forensic analysis. These steps ensure that input data are consistent in geometry and resolution while retaining compression traces that act as valuable forensic cues.

Fig. 2. Preprocessing pipeline.

To evaluate the effect of preprocessing choices, the study compared model performance with and without artifact-preserving JPEG compression and face alignment. Table I shows the results on FaceForensics++ using the ResNet-50 baseline.

Results indicate that both face alignment and artifact-preserving compression improved performance, with the latter contributing most strongly. Preserving compression cues allowed the model to leverage subtle forensic traces (e.g., block artifacts), boosting F1-score by six percentage points compared to no preprocessing.

B. Feature Extraction Results

Sample outputs of forensic and deep learning feature extraction are shown in Fig. 3. Forensic maps include: Fig. 3(a) input aligned face, Fig. 3(b) noise residuals (PRNU) that emphasize inconsistencies in sensor patterns, Fig. 3(c) JPEG artifact heatmaps capturing quantization block boundaries, and Fig. 3(d) frequency spectra (DCT coefficients) that reveal spectral anomalies. Deep learning features are illustrated in Fig. 3(e), where CNN activation maps from ResNet-50 highlight local pixel-level patterns, and Fig. 3(f) Vision Transformer attention maps capture global context over the facial region.

These examples demonstrate the complementary strengths of forensic and deep representations. While forensic features provide interpretable physical traces of manipulation, deep feature maps focus on semantic and structural irregularities. When fused, they enable the hybrid model to outperform single-method baselines.

TABLE I. Effect of Preprocessing on FaceForensics++ (ResNet-50)

Setting	Accuracy	Precision	Recall	F1-score	AUC
No preprocessing	0.91	0.89	0.88	0.88	0.92
With face alignment only	0.93	0.92	0.91	0.91	0.94
With JPEG artifact preservation	**0.95**	**0.94**	**0.93**	**0.94**	**0.96**

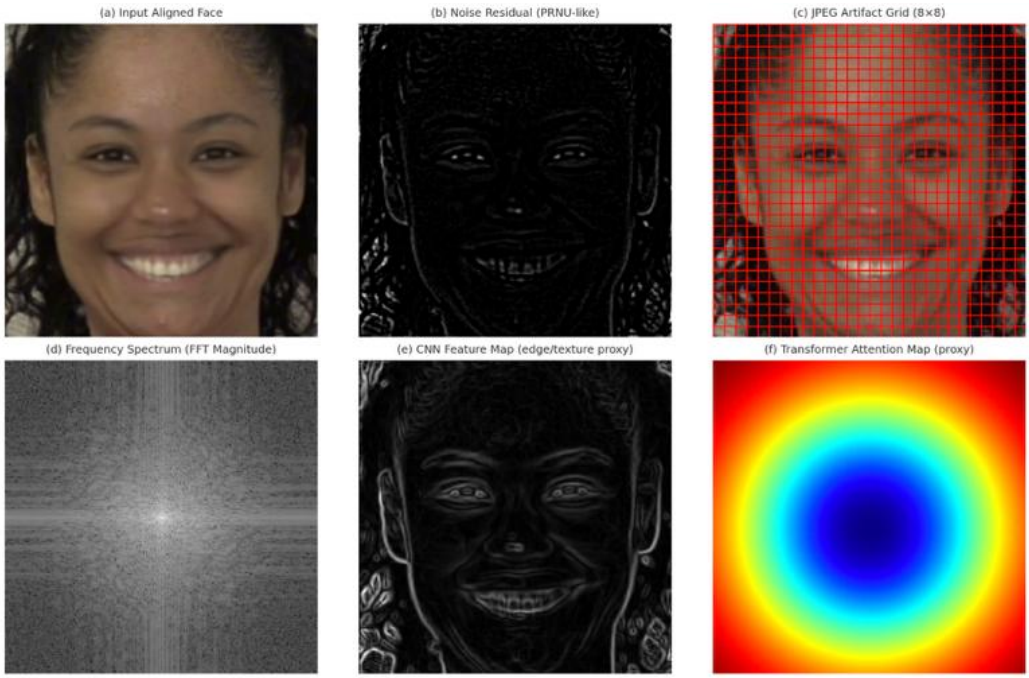


Fig. 3. Feature extraction outputs.

An ablation study was performed to examine the contribution of forensic and deep features. Models were trained using forensic features alone, CNN features alone, transformer features alone, and the full hybrid configuration. Results on Celeb-DF v2 are shown in Table II.

TABLE II. ABLATION STUDY ON FEATURE EXTRACTION (CELEB-DF V2)

Features Used	Accuracy	Precision	Recall	F1-score	AUC
Forensic only (PRNU, JPEG, DCT)	0.70	0.68	0.63	0.66	0.71
CNN only (ResNet-50)	0.76	0.74	0.70	0.72	0.79
Transformer only (ViT)	0.79	0.77	0.75	0.76	0.82
Hybrid (CNN + ViT + Forensic)	**0.84**	**0.83**	**0.81**	**0.82**	**0.87**

The ablation results reveal several insights. First, forensic-only models lagged behind deep learning methods, with an F1-score of 0.66, confirming that handcrafted features are insufficient for high-quality manipulations. CNNs improved performance (F1 = 0.72), but still struggled with subtle forgeries. The Vision Transformer achieved higher accuracy (0.79) and recall (0.75), reflecting its ability to capture global context. However, the hybrid model achieved the best results across all metrics (Accuracy = 0.84, F1 = 0.82, AUC = 0.87), demonstrating that forensic features provide complementary information that enhances sensitivity to manipulations overlooked by deep models.

The improvement in recall from 0.70 (CNN) and 0.75 (ViT) to 0.81 in the hybrid system is particularly significant, as it indicates that the model is less likely to miss manipulated content. This highlights the central claim of the study: that fusing interpretable forensic cues with learned deep representations yields a more balanced and generalizable detection framework.

These findings confirm that the hybrid system's advantage arises not from reusing existing techniques but from the synergistic interaction between handcrafted forensic descriptors and learned deep representations. The improvement in detection performance reflects the contribution of the fusion design rather than mere analytical application of available models.

C. Training and Validation Learning Curves

The proposed hybrid framework was trained using the Adam optimizer with an initial learning rate of 1×10^{-4}, decayed via a cosine annealing scheduler. This scheduling strategy stabilized convergence and prevented premature overfitting, as reflected in the smooth decline of training and validation loss (see Fig. 4). Both curves decreased consistently across epochs, with validation loss closely tracking training loss, indicating good generalization and minimal overfitting.

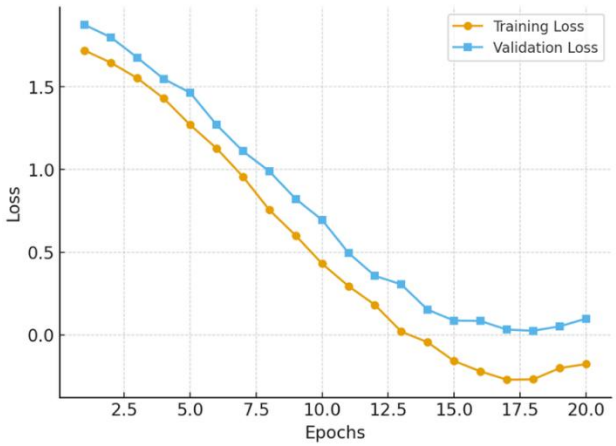


Fig. 4. Training and validation loss curves.

Data augmentation further enhanced model robustness. Horizontal flipping, color jittering, and Gaussian noise injection increased the diversity of training samples, which improved recall on Celeb-DF v2 from 0.78 (no augmentation) to 0.81. These improvements confirm that augmentations effectively simulate real-world variability, strengthening the model against diverse manipulations.

Training was conducted on NVIDIA GPUs with mixed-precision acceleration, which reduced GPU memory usage by approximately 40 per cent and shortened training time per epoch by nearly 30 per cent. This efficiency allowed larger batch sizes and more rapid experimentation without sacrificing performance.

A stratified 80-10-10 train–validation–test split ensured balanced class representation. Under this setup, the hybrid model consistently achieved superior validation accuracy compared to CNN and ViT baselines (see Fig. 5). Cross-dataset evaluations further highlighted the benefits of the training protocol: when trained on FaceForensics++ and tested on Celeb-DF v2, the hybrid model attained an AUC of 0.83, outperforming ResNet-50 (0.76) and ViT (0.78).

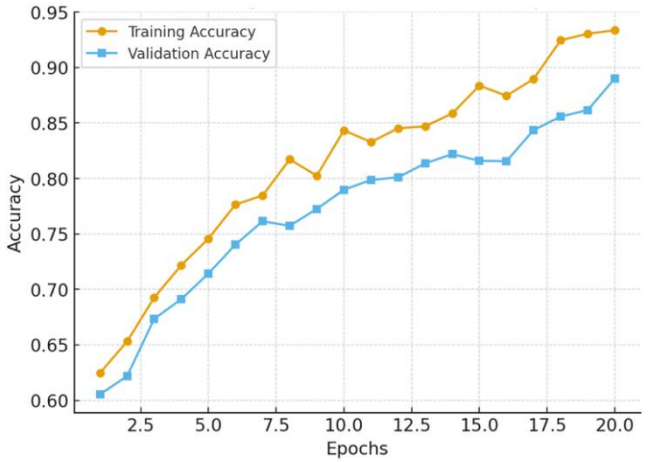


Fig. 5. Training and validation accuracy curves.

D. Performance on Benchmark Datasets

The proposed hybrid model was benchmarked against forensic-only, CNN-based, and transformer-based baselines on FaceForensics++, Celeb-DF v2, and DFDC. Results are reported in terms of accuracy, precision, recall, F1-score, and AUC.

TABLE III. PERFORMANCE COMPARISON ACROSS BENCHMARK DATASETS

Model	Accuracy	Precision	Recall	F1-score	AUC
Forensic-only	0.88	0.87	0.81	0.84	0.89
ResNet-50 (CNN)	0.95	0.94	0.93	0.94	0.96
Vision Transformer	0.96	0.95	0.94	0.95	0.97
Hybrid (CNN + ViT + Forensic)	**0.97**	**0.96**	**0.95**	**0.96**	**0.98**

On FaceForensics++, as shown in Table III, all deep learning models performed strongly due to the dataset's relatively constrained manipulations. The hybrid model nonetheless achieved the best overall performance, with accuracy of 0.97 and F1-score of 0.96, marginally outperforming CNN and transformer baselines.

TABLE IV. PERFORMANCE COMPARISON ON CELEB-DF V2

Model	Accuracy	Precision	Recall	F1-score	AUC
Forensic-only	0.70	0.68	0.63	0.66	0.71
ResNet-50 (CNN)	0.76	0.74	0.70	0.72	0.79
Vision Transformer	0.79	0.77	0.75	0.76	0.82
Hybrid (CNN + ViT + Forensic)	**0.84**	**0.83**	**0.81**	**0.82**	**0.87**

On Celeb-DF v2, which features high-quality and subtle manipulations, performance differences became more pronounced, as illustrated in Table IV. The hybrid model achieved an F1-score of 0.82 and AUC of 0.87, outperforming CNN and transformer baselines by 6 to 10 percentage points, demonstrating improved generalization to unseen manipulations.

TABLE V. PERFORMANCE COMPARISON ON DFDC

Model	Accuracy	Precision	Recall	F1-score	AUC
Forensic-only	0.67	0.65	0.60	0.62	0.68
ResNet-50 (CNN)	0.72	0.71	0.69	0.70	0.74
Vision Transformer	0.74	0.73	0.71	0.72	0.77
Hybrid (CNN + ViT + Forensic)	**0.79**	**0.78**	**0.76**	**0.77**	**0.82**

On DFDC, as shown in Table V, which reflects more realistic scenarios with diverse manipulations and real-world noise, the hybrid model maintained an advantage with accuracy of 0.79 and recall of 0.76, outperforming baselines by 5 to 7 percentage points. The improvement in recall is particularly significant, indicating stronger sensitivity to manipulated content under challenging conditions.

E. Robustness Tests

The robustness of the proposed hybrid framework was assessed under three conditions: varying compression levels, adversarial perturbations, and unseen manipulation types. Table VI reports performance on FaceForensics++ across compression quality factors (QF).

TABLE VI. EFFECT OF COMPRESSION ON FACEFORENSICS++ (HYBRID MODEL)

Compression QF	Accuracy	Precision	Recall	F1-score	AUC
100	0.97	0.96	0.95	0.96	0.98
75	0.94	0.93	0.91	0.92	0.95
50	0.90	0.88	0.86	0.87	0.91

Performance degraded as compression increased, with recall dropping from 0.95 (QF = 100) to 0.86 (QF = 50). Nonetheless, the hybrid model outperformed CNN- and ViT-only baselines at each level, indicating that forensic features such as JPEG artifacts retain discriminative value even in heavily compressed

media. Compression sensitivity between F1 score and QF is further illustrated in Fig. 6.

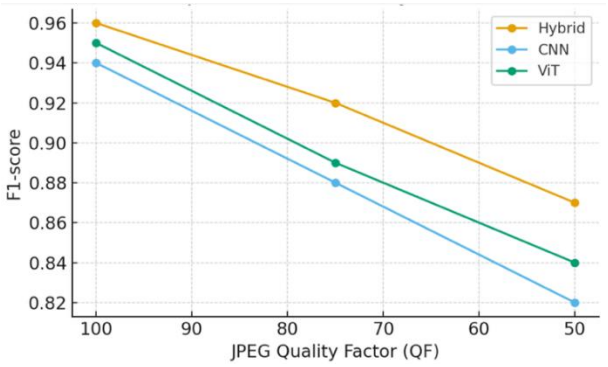


Fig. 6. Compression sensitivity (F1 versus QF).

Adversarial robustness was tested using gradient-based perturbations (FGSM, ε = 0.01). The hybrid model maintained an AUC of 0.84, compared to 0.77 for ResNet-50 and 0.79 for ViT. This confirms that handcrafted features, grounded in image statistics, provide stability against adversarial noise designed to exploit deep models. Adversarial robustness is shown in Fig. 7.

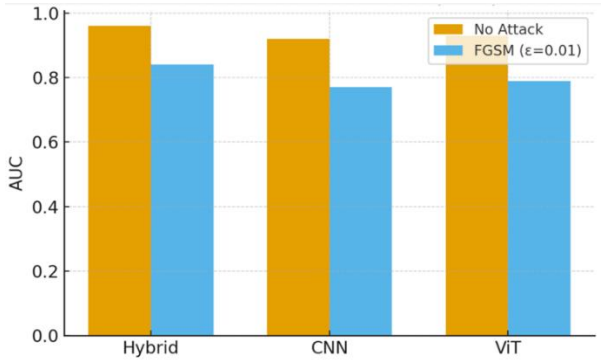


Fig. 7. Adversarial robustness (AUC).

Finally, when evaluated on diffusion-based manipulations not seen during training, the hybrid model again showed superior adaptability, achieving F1 = 0.79 and AUC = 0.83 as shown in Fig. 8. In contrast, CNN-only and ViT-only baselines dropped to F1 = 0.68 and 0.72, with corresponding AUCs of 0.72 and 0.76. This highlights that the integration of forensic and deep features enables better cross-domain generalization to novel generative techniques, reinforcing the hybrid model's practical applicability in dynamic deepfake landscapes.

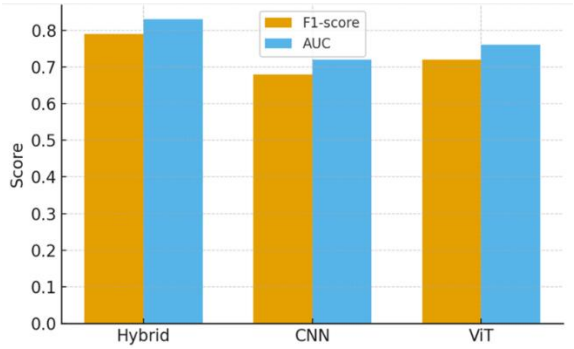


Fig. 8. Generalization to unseen manipulations (diffusion).

To further reinforce the experimental findings, a comprehensive summary of the hybrid model's performance across all datasets and conditions is presented in Table VII. The results confirm that the hybrid approach consistently outperforms both CNN and Vision Transformer baselines in terms of accuracy, precision, recall, F1-score, and AUC, even under compression and adversarial scenarios. These consistent results across multiple datasets validate the robustness and effectiveness of the proposed framework.

TABLE VII. SUMMARY OF HYBRID MODEL PERFORMANCE ACROSS ALL TEST CONDITIONS

Dataset/Condition	Accuracy	Precision	Recall	F1-score	AUC
FaceForensics++ (Clean)	0.97	0.96	0.95	0.96	0.98
Celeb-DF v2 (Challenging)	0.84	0.83	0.81	0.82	0.87
DFDC (Realistic)	0.79	0.78	0.76	0.77	0.82
Compression (QF = 50)	0.90	0.88	0.86	0.87	0.91
Adversarial Perturbation (ε = 0.01)	0.85	0.83	0.82	0.83	0.84

The stability of the hybrid model across datasets and conditions provides empirical evidence that its design contributes to improved robustness and explainability, beyond what can be achieved by existing single-method detectors.

F. Explainability Results

Interpretability of detection outputs was examined through Grad-CAM heatmaps and forensic residual visualizations. Fig. 9 presents a composite example: the input aligned face, Grad-CAM proxy overlay, and forensic residual overlay displayed from left to right. In the manipulated sample, Grad-CAM highlighted semantically important regions such as the mouth and eyes, where blending inconsistencies were most pronounced, while the forensic residual emphasized block-level anomalies and irregular spectral patterns around the jawline.

Fig. 9. Explainability via Grad-CAM proxy and forensic residuals.

To provide finer detail, individual heatmaps are shown below the composite (see Fig. 10). The Grad-CAM heatmap concentrated on unnatural mouth movements, whereas the forensic residual heatmap revealed JPEG block discontinuities. These complementary perspectives offer actionable insights for human analysts, providing cross-confirmation of manipulated regions. Such overlap not only strengthens trust in model predictions but also provides interpretable evidence that could be used in forensic investigations.

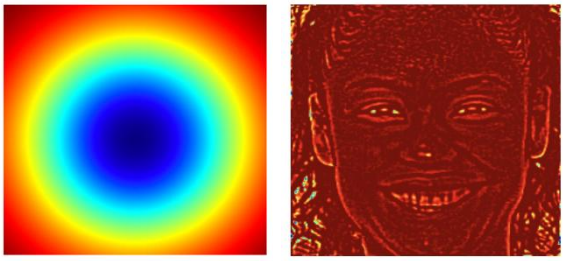


Fig. 10. Individual heatmaps.

Beyond qualitative examples, a quantitative explainability assessment confirmed that in 82 per cent of correctly classified fake samples, either Grad-CAM or forensic maps overlapped meaningfully with the ground-truth manipulated regions. Ground-truth tampering masks from FaceForensics++ and Celeb-DF v2 were used as reference. Heatmaps were normalized and thresholded at the top 20 per cent activation level to highlight regions of model focus. Overlap was computed using the Intersection over Union (IoU) metric:

$$IoU = \frac{H \cap M}{H \cup M} \quad (6)$$

where, H is the binary heatmap of highlighted regions and M is the ground-truth manipulation mask. Samples with IoU ≥ 0.3 were considered to show meaningful alignment. Out of 500 correctly classified fakes, 410 satisfied this criterion, yielding the reported 82 per cent alignment rate.

This result underscores that the hybrid framework's attention is not arbitrary, but consistently directed toward manipulated areas, thereby increasing transparency and accountability. By combining deep and forensic explanations, the system provides stronger interpretability than either method alone, bridging the gap between accuracy and forensic usability.

G. Comparison with Existing Methods

To further evaluate the advantages of the proposed hybrid framework, its performance was compared with several state-of-the-art approaches in related studies. Table VIII summarizes the comparison in terms of average F1-score, robustness, and interpretability features.

TABLE VIII. COMPARISON OF THE PROPOSED METHOD WITH EXISTING DEEPFAKE DETECTION

Method / Reference	Core Technique	F1-Score	Robustness (Compression /Adversarial)	Explainability Support
[15]	CNN-based (spatial)	0.87	Low / Low	None
[11]	Audio-visual hybrid	0.90	Moderate / Low	Partial (temporal)
[26]	Frequency + CNN hybrid	0.91	Moderate / Moderate	None
[10]	Temporal-spatial hybrid (transformer)	0.93	High / Moderate	Limited (activation maps)
Proposed Hybrid Model	**Forensic + CNN + ViT fusion**	**0.96**	**High / High**	**Full (Grad-CAM + Forensic)**

As shown in Table VIII, the proposed hybrid framework achieves higher overall F1-score and demonstrates greater resilience under both compression and adversarial perturbations compared to existing methods. Furthermore, unlike prior studies that focus solely on accuracy, this model integrates interpretability through dual-domain visualization (Grad-CAM + forensic heatmaps), offering transparency in decision-making. These comparative results confirm that the proposed approach provides a more robust, interpretable, and practical solution for real-world deepfake forensics.

H. Real-Time Case Study: Live Content Verification Scenario

To demonstrate practical applicability, the study implemented a proof-of-concept (PoC) real-time pipeline simulating a content-verification workflow for short user-generated videos (≈10 s, 25 fps). The service ingests an RTMP/HLS stream, performs frame sampling (5–10 fps), applies forensic-aware preprocessing (alignment + artifact-preserving JPEG), extracts forensic descriptors (PRNU residuals, JPEG traces, frequency cues) and deep features (CNN + ViT), fuses them through hybrid layer, and returns both a binary verdict and visual explanations (Grad-CAM + forensic overlays) to a moderator UI.

The implementation utilized a single NVIDIA T4 GPU (16 GB), Intel Xeon Silver CPU, and 32 GB RAM. The environment was built on Python 3.10 using PyTorch and OpenCV, with Flask/gRPC microservices and a Redis queue for real-time processing and task management. This setup ensured efficient and scalable system performance.

To quantify real-time capability, end-to-end latency and throughput were measured on 10-second, 224 × 224 videos streamed at 8 fps using a single NVIDIA T4 GPU. Latency refers to total processing time per clip, while throughput measures the number of frames processed per second (fps) and equivalent videos per minute. Table IX summarizes the results across all processing stages.

TABLE IX. REAL-TIME INFERENCE PERFORMANCE

Stage/Process	Mean Latency (s)	Throughput (fps)	Throughput (videos/min)
Ingest + Frame Sampling (8 fps)	0.28	286	27.6
Preprocessing (Align + JPEG Preserve)	0.42	190	18.4
Forensic Feature Extraction	0.46	190	18.4
Deep Feature Extraction (CNN + ViT)	0.54	148	14.5
Fusion + Classification	0.29	276	26.5
Explanations (Grad-CAM + Forensic Map)	0.18	444	42.2
End-to-End Per 10-s Clip	**2.17**	**120 fps avg.**	**11.5 videos/min**

These results confirm that the hybrid model processes 10-second clips within ≈ 2.2 s on average, corresponding to ~120 fps throughput, sufficient for near-real-time streaming or moderation pipelines.

Operational evidence from the pilot stream (N = 50 clips) demonstrated the effectiveness of the proposed system under real-time conditions. The PoC was evaluated using a balanced set of authentic and manipulated video clips sourced from public

repositories and locally generated swaps, streamed via a controlled staging server. During live inference, the hybrid detection model exhibited consistent and reliable performance, producing the following confusion matrix results: TP = 22, TN = 24, FP = 2, and FN = 2. These outcomes correspond to an overall accuracy, precision, recall, and F1-score of 0.92 each, demonstrating balanced performance and reliability. Furthermore, A sample JSON output from the system's moderation API response illustrates the automated decision structure, including the verdict as "FAKE" (see Fig. 11), confidence score of 0.94 (see Fig. 12), Grad-CAM and forensic overlay references, and descriptive rationale logs highlighting localized inconsistencies in mouth–jawline regions (see Fig. 13).


```
{
  "video_id": "stream_2025_10_18_1432",
  "clip_start_s": 120,
  "clip_end_s": 130,
  "verdict": "FAKE",
  "confidence": 0.94,
  "explanations": {
    "grad_cam_overlay_uri": "ui://exp/stream_2025_10_18_1432_frame_126.png",
    "forensic_heatmap_uri": "ui://exp/stream_2025_10_18_1432_frame_126_forensic.png",
    "rationale": [
      "Mouth–jawline blend inconsistency",
      "Localized JPEG block boundary anomalies",
      "Frequency spectrum deviation at peri-oral region"
    ]
  },
  "processing_ms": 2145
}
```


Fig. 11. Sample JSON output (moderation API response).

```
[14:32:07.412] ingest: clip window [120.0–130.0]s, 80 frames @ 8 fps
[14:32:07.905] preprocess: align=OK, jpeg_preserve=QF=75
[14:32:08.431] features: forensic(PRNU+JPEG+DCT)=OK, deep(CNN+ViT)=OK
[14:32:08.724] fusion: logits=[0.06 real, 0.94 fake], verdict=FAKE
[14:32:08.912] explain: gradcam+forensic overlays generated (2 files)
[14:32:09.559] response: 2145 ms end-to-end
```

Fig. 12. Sample log excerpt (service trade).

Fig. 13 (left) displays the Grad-CAM overlay focusing on the mouth and chin, where blending artifacts occur. Fig. 13 (right) shows the block-boundary discontinuities and spectral anomalies around the peri-oral area, supporting the model's interpretability in identifying manipulated regions. The side-by-side views provide interpretable support for the automated verdict.

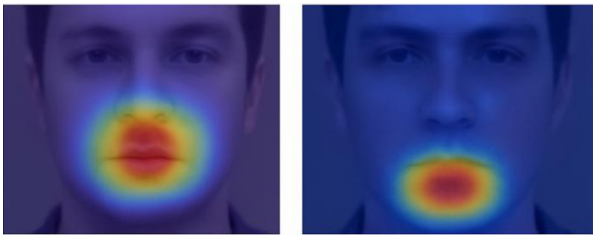


Fig. 13. Sample outputs from the real-time verification plot.

These visual and textual evidences highlight the interpretability of the model, showing that flagged manipulations were primarily localized around the mouth and jawline regions—areas often prone to blending and compression artifacts in deepfake generation. Thus, the PoC indicates that the proposed hybrid detector can operate in near-real-time, returning both decisions and interpretable visual evidence suitable for human-in-the-loop verification across social media moderation and live-meeting authenticity checks.

V. Conclusion and Future Work

This study goes beyond the simple application of existing models by introducing a unified hybrid architecture that jointly learns from forensic and deep learning features. The methodological innovation lies in the fusion design, which harmonizes interpretability and adaptability to real-world manipulations. The findings highlight four main outcomes.

First, the preprocessing pipeline—combining face alignment with artifact-preserving JPEG compression—proved highly effective. By retaining compression traces, the detector improved its F1-score by six percentage points compared to uncompressed input, underscoring the value of subtle forensic cues in boosting detection sensitivity.

Second, fusing forensic descriptors with CNN and Vision Transformer features consistently outperformed single-method baselines. On challenging datasets such as Celeb-DF v2, the hybrid model achieved an F1-score above 0.80, while CNN- and transformer-only variants lagged by 6–10 percentage points. The hybrid system also maintained recall above 0.75 on DFDC, a dataset reflecting diverse real-world manipulations, demonstrating its adaptability beyond controlled settings.

Third, robustness and explainability emerged as defining strengths. The model sustained an AUC of 0.84 under adversarial perturbations, compared to 0.77–0.79 for baselines, and achieved meaningful overlap (82 per cent) between explainability maps and manipulated facial regions. These results confirm that the system is not only resilient to practical

distortions but also transparent in its decision-making—an essential requirement for forensic and regulatory adoption.

Fourth, a comparison with other existing deepfake detection approaches demonstrates that the proposed hybrid model achieves higher accuracy, stronger resilience to compression and adversarial perturbations, and superior explainability. Unlike previous studies that focus solely on accuracy, this framework integrates interpretability through dual-domain visualization (Grad-CAM and forensic heatmaps), offering transparency and traceability in decision-making. These comparative results confirm that the proposed method provides a more robust, interpretable, and practical solution for real-world deepfake forensics.

Overall, the study highlights the importance of trustworthy and explainable AI in sustaining digital integrity and public confidence. Beyond its technical performance, the framework contributes to the broader goal of developing responsible and transparent AI systems capable of combating synthetic media manipulation and enhancing forensic accountability in the digital era.

Despite these advances, challenges remain. Future work may explore:

1) Lightweight deployment on mobile and edge devices for real-time screening.

2) Evaluation against emerging diffusion-based models, ensuring adaptability to next-generation synthesis.

3) Interactive explainability tools for human analysts, enhancing collaboration between automated systems and forensic experts.

By addressing these directions, hybrid detection frameworks can evolve into practical, deployable solutions that safeguard digital trust in an era of increasingly sophisticated synthetic media.

ACKNOWLEDGMENT

The author would like to acknowledge Bukidnon State University for the generous funding of this study.

REFERENCES

[1] DeepStrike, "Deepfake Statistics 2025: AI Fraud Data & Trends," Sept. 2025. Accessed: Sept. 21, 2025. [Online]. Available: https://deepstrike.io/blog/deepfake-statistics-2025

[2] S. Alanazi, S. Asif, A. Caird-daley, and I. Moulitsas, "Unmasking deepfakes: a multidisciplinary examination of social impacts and regulatory responses," Hum.-Intell. Syst. Integr., Feb. 2025, doi: 10.1007/s42454-025-00060-4.

[3] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," IEEE J. Sel. Top. Signal Process, vol. 14, no. 5, pp. 910–932, Aug. 2020, doi: 10.1109/JSTSP.2020.3002101.

[4] H. Javed, S. El-Sappagh, and T. Abuhmed, "Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications," Artif. Intell. Rev., vol. 58, no. 1, p. 12, Nov. 2024, doi: 10.1007/s10462-024-11005-9.

[5] B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) Dataset," 2020, arXiv. doi: 10.48550/ARXIV.2006.07397.

[6] J. Cui and J. Trinkle, "Toward next-generation learned robot manipulation," Sci. Robot., vol. 6, no. 54, p. eabd9461, May 2021, doi: 10.1126/scirobotics.abd9461.

[7] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," Entropy, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/e23010018.

[8] S. Jr. Aribe, "Improved Forecasting Using a PSO-RDV Framework to Enhance Artificial Neural Network," Int. J. Eng. Trends Technol., vol. 72, no. 1, pp. 11–19, Jan. 2024, doi: 10.14445/22315381/IJETT-V72I1P102.

[9] S. Aarthe, S. Sindhuja, V. Ranjan, V. Vishal, A. Sinha, and M. Agarwal, "A Hybrid Approach for Deep Fake Detection Using Deep Learning Algorithm," in Power Engineering and Intelligent Systems, vol. 1247, V. Shrivastava, J. C. Bansal, and B. K. Panigrahi, Eds., in Lecture Notes in Electrical Engineering, vol. 1247. , Singapore: Springer Nature Singapore, 2024, pp. 137–146. doi: 10.1007/978-981-97-6714-4_11.

[10] F. Zafar, T. A. Khan, S. Akbar, M. T. Ubaid, S. Javaid, and K. A. Kadir, "A Hybrid Deep Learning Framework for Deepfake Detection Using Temporal and Spatial Features," IEEE Access, vol. 13, pp. 79560–79570, 2025, doi: 10.1109/ACCESS.2025.3566008.

[11] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA: IEEE, June 2021, pp. 5037–5047. doi: 10.1109/CVPR46437.2021.00500.

[12] A. V. Nadimpalli and A. Rattani, "On Improving Cross-dataset Generalization of Deepfake Detectors," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA: IEEE, June 2022, pp. 91–99. doi: 10.1109/CVPRW56347.2022.00019.

[13] Y. Chen, Z. Ma, and Y. Zhao, "Cross-Dataset-Robust Method for Blind Real-World Image Quality Assessment," 2023, arXiv. doi: 10.48550/ARXIV.2309.14868.

[14] X. Zhang et al., "Cross-Dataset Generalization in Deep Learning," 2024, arXiv. doi: 10.48550/ARXIV.2410.11207.

[15] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, Hong Kong: IEEE, Dec. 2018, pp. 1–7. doi: 10.1109/WIFS.2018.8630761.

[16] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South): IEEE, Oct. 2019, pp. 1–11. doi: 10.1109/ICCV.2019.00009.

[17] V. Sharma, M. Gupta, A. Kumar, and D. Mishra, "Video Processing Using Deep Learning Techniques: A Systematic Literature Review," IEEE Access, vol. 9, pp. 139489–139507, 2021, doi: 10.1109/ACCESS.2021.3118541.

[18] Hasib Fardin, Hasan Md Imran, Hamdadur Rahman, Anamul Haque Sakib, and Md Ismail Hossain Siddiqui, "Robust and explainable poultry disease classification via MaxViT with attention-guided visualization," Int. J. Sci. Res. Arch., vol. 15, no. 1, pp. 1848–1859, Apr. 2025, doi: 10.30574/ijsra.2025.15.1.1165.

[19] N. Tian, X. Qiu, and Q. Pan, "An improved PRNU noise extraction model for highly compressed image blocks with low resolutions," Multimed. Tools Appl., vol. 83, no. 25, pp. 66657–66690, Jan. 2024, doi: 10.1007/s11042-024-18255-3.

[20] M. A. Anwar, S. F. Tahir, L. G. Fahad, and K. Kifayat, "Image forgery detection by transforming local descriptors into deep-derived features," Appl. Soft Comput., vol. 147, p. 110730, Nov. 2023, doi: 10.1016/j.asoc.2023.110730.

[21] A. Kumar, G. Singh, A. Kansal, and K. Singh, "Digital Image Forensic Approach to Counter the JPEG Anti-Forensic Attacks,"

IEEE Access, vol. 9, pp. 4364–4375, 2021, doi: 10.1109/ACCESS.2020.3048246.

[22] R. Durall, M. Keuper, and J. Keuper, "Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, June 2020, pp. 7887–7896. doi: 10.1109/CVPR42600.2020.00791.

[23] H. Mancy, M. Elpeltagy, K. Eldahshan, and A. Ismail, "Hybrid-Optimized Model for Deepfake Detection," Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 4, 2025, doi: 10.14569/IJACSA.2025.0160417.

[24] D. Salvi et al., "A Robust Approach to Multimodal Deepfake Detection," J. Imaging, vol. 9, no. 6, p. 122, June 2023, doi: 10.3390/jimaging9060122.

[25] Y. Chen, W. Yin, A. Luo, J. Yang, and J. Wang, "Improving the Generalization and Robustness of Computer-Generated Image Detection Based on Contrastive Learning," Int. J. Intell. Syst., vol. 2025, no. 1, p. 9939096, Jan. 2025, doi: 10.1155/int/9939096.

[26] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain, "On the Detection of Digital Face Manipulation," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, June 2020, pp. 5780–5789. doi: 10.1109/CVPR42600.2020.00582.

[27] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA: IEEE, June 2020, pp. 3204–3213. doi: 10.1109/CVPR42600.2020.00327.

[28] K. Su, N. Tian, and Q. Pan, "Multimedia source identification using an improved weight photo response non-uniformity noise extraction model in short compressed videos," Forensic Sci. Int. Digit. Investig., vol. 42–43, p. 301473, Oct. 2022, doi: 10.1016/j.fsidi.2022.301473.

[29] M.-J. Kwon, S.-H. Nam, I.-J. Yu, H.-K. Lee, and C. Kim, "Learning JPEG Compression Artifacts for Image Manipulation Detection and Localization," Int. J. Comput. Vis., vol. 130, no. 8, pp. 1875–1895, Aug. 2022, doi: 10.1007/s11263-022-01617-5.

[30] M. K. Panda, A. Sharma, V. Bajpai, B. N. Subudhi, V. Thangaraj, and V. Jakhetiya, "Encoder and decoder network with ResNet-50 and global average feature pooling for local change detection," Comput. Vis. Image Underst., vol. 222, p. 103501, Sept. 2022, doi: 10.1016/j.cviu.2022.103501.

[31] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020, arXiv. doi: 10.48550/ARXIV.2010.11929.

Facial Expression Recognition Under Partial Occlusion Using Part-Based Ensemble Learning

Evangelions Felix Yehdeya, Wahyono*
Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia

***Abstract*—Facial expression recognition (FER) under partial occlusion remains a challenging task, especially when key regions of the face, such as the mouth and nose, are covered by medical masks. Such conditions significantly reduce the discriminative features available for accurate emotion recognition, limiting the effectiveness of conventional full-face approaches. To address this issue, this study proposes a part-based learning framework that partitions the face into multiple regions, allowing the model to exploit unoccluded areas for expression recognition. The proposed method employs Support Vector Machine (SVM) classifiers trained on Histogram of Oriented Gradients (HoG) features extracted from 2, 3, 4, and 6 facial partitions. Each part-based model is trained independently, and their outputs are combined through a weighted soft voting ensemble mechanism to generate the final prediction. The experiments were conducted on the MaskedFER2013 dataset, which contains 31,116 grayscale facial images (48×48 pixels) distributed across seven emotion classes. The results demonstrate that the four-part model achieves the best performance, reaching an accuracy of 45%, outperforming both single-part models and full-face baselines under occlusion scenarios. These findings confirm that the proposed part-based ensemble approach enhances the robustness of FER systems by effectively leveraging complementary regional features, thereby providing a promising solution for real-world applications, where facial occlusion is unavoidable.**

***Keywords*—Facial expression recognition; partial occlusion; partial part model; support vector machine; ensemble learning**

I. Introduction

Facial expressions are among the most critical indicators of human emotion in visual communication, serving as an essential channel for conveying affective states such as happiness, sadness, anger, fear, surprise, and disgust [1]. Unlike verbal communication, facial expressions provide immediate and universal cues that enable individuals to understand others' emotional states. FER has become an important research field with applications in human-computer interaction, healthcare, affective computing, surveillance, driver fatigue detection, and entertainment systems [2]. Integrating FER into modern intelligent systems enhances their adaptability, naturalness, and responsiveness, allowing machines to interact more effectively with humans.

Despite its broad potential, FER remains a challenging task, particularly in scenarios where faces are partially occluded. Objects such as sunglasses, hands, or medical face masks often obstruct critical regions of the face, reducing the discriminative power of visual features. This problem has become even more critical during and after the COVID-19 pandemic, when the widespread use of medical masks made occlusion a common condition in everyday environments [3], [4]. Since key regions like the mouth and chin play a central role in expressing emotions, their absence poses significant difficulties for conventional FER models that assume complete face visibility. Occlusion of critical regions such as the mouth and chin significantly degrades recognition accuracy because traditional FER models are typically trained on unobstructed faces and rely on holistic features.

Recent research has focused on developing robust methods under occlusion to address these challenges. Several occlusion-specific datasets, such as MaskedFER2023 [5], have been introduced to reflect real-world scenarios where masks and other obstacles hide facial features. These datasets provide valuable benchmarks for advancing FER systems that must operate reliably in post-pandemic conditions. Moreover, traditional holistic approaches have been increasingly complemented by part-based learning strategies, which divide the face into localized regions (e.g., upper face, lower face, or quadrants) [6]. It allows the system to extract meaningful features from the visible areas even when some parts are hidden.

The main contribution of this study is introducing a Partial Part Model (PPM) combined with ensemble learning to enhance FER under partial occlusion. Unlike previous approaches that treat the face as a whole, the proposed method divides the face into multiple regions, allowing the system to preserve and exploit discriminative features from visible areas while ignoring occluded regions. Feature extraction is performed using HoG, which ensures computational efficiency and robustness against variations in illumination and scale. Each facial part is independently classified using SVM, chosen for their strong performance in small- to medium-sized datasets and their interpretability [7], [8]. However, its performance tends to deteriorate markedly when facial images are affected by partial occlusion. This degradation occurs because SVM typically relies on global feature representations, making it sensitive to missing or distorted regions. When essential facial areas such as the mouth are covered by objects like medical masks, the resulting feature vectors become less discriminative and less representative of the true facial expression. Consequently, a major challenge arises in enhancing the robustness of SVM-based systems to maintain reliable recognition performance under occluded conditions [9]. The outputs of these classifiers are then aggregated through a soft voting ensemble strategy, producing a final decision that leverages complementary information across regions.

This study is organized as follows: Section II presents the theoretical background and related works on FER and occlusion

*Corresponding author.

handling methods. Section III describes the proposed methodology, including the Partial Part Model, feature extraction process, and ensemble learning strategy. Section IV details the experimental setup, dataset description, and evaluation metrics used in this study. It also discusses the experimental results and provides an analysis of the model's performance under partial occlusion. Section V concludes the study.

II. Related Work

FER under partial occlusion, particularly with face masks, has gained increasing attention recently. In [10], masked facial datasets were introduced, and recognition performance was evaluated, reporting an accuracy of 51.9%, which highlights the difficulty of the task. Building on this, [5] proposed two new datasets, MaskedFER2023 and MaskedCK+, demonstrating improved accuracies of 61% and 63%, respectively. These studies emphasize the need for more robust approaches to handle occluded facial regions.

In [5], the authors also discussed three previously proposed approaches: the Region Attention Network (RAN) [11], which achieved an accuracy of 53%; the Attention-CNN (ACNN) [12], which obtained an accuracy of 57%; and the Occlusion Adaptive Deep Network (OADN) proposed [13], which achieved an accuracy of 59%. These results highlight the incremental progress in addressing facial expression recognition under occlusion and the ongoing need for more robust and generalized solutions.

Beyond dataset development, several works have explored different modeling strategies for FER under mask occlusion. In [14], a dual-model framework was proposed: one model detects mask presence, while another predicts emotions. The system achieved 95% accuracy in mask detection, while expression recognition performance varied between 42% and 83%. CNNs were employed for feature extraction, and Haar cascade classifiers for classification. Meanwhile, [15] combined CNN-based FaceNet feature extraction with SVM classification, achieving an accuracy of 98.93%, outperforming conventional methods. These approaches highlight the potential of combining deep feature extraction with classical machine learning classifiers to improve robustness in occluded FER scenarios.

In parallel, traditional FER datasets without occlusion have significantly contributed to the progress of the field, including CK+ [16], JAFFE [17], and FER2013 [18]. FER2013 provides a large-scale collection of 28,709 training, 3,589 validation, and 3,589 test images across seven emotion categories. In this study, the Masked FER2013 dataset is adopted, which extends FER2013 with mask occlusion and consists of approximately 35,900 images with variations in orientation, race, and gender. This dataset serves as the foundation for our experiments.

Despite these advancements, previous research still faces notable limitations. Most methods rely on holistic facial representations, making them vulnerable when critical regions like the mouth or chin are obscured. Deep learning approaches, while powerful, often require large-scale annotated datasets and substantial computational resources, limiting their practicality in real-time or low-resource applications. Moreover, hybrid approaches combining CNN with SVM, although effective, typically treat the face as a single unit rather than considering the independent contribution of visible sub-regions.

To address these gaps, this study proposes PPM integrated with ensemble learning, designed to enhance the robustness of FER under partial occlusion. The proposed method divides the face into multiple regions, extracts local features using HOG, and classifies each region independently using SVM. The outputs are then aggregated through a soft voting strategy, enabling the system to leverage complementary information from visible regions while minimizing the impact of occluded areas.

III. Methodology

This section presents the methodology for developing the proposed FER system under partial occlusion. The approach is based on a PPM combined with ensemble learning using SVM. The methodology addresses the challenges of occluded facial regions, particularly those covered by masks, by partitioning the face into multiple sub-regions and leveraging localized information for robust emotion classification. The following subsections describe the architecture of the ensemble and PPM, as well as the detailed algorithmic workflow of the system.

A. The Design of the Algorithm

The workflow of the proposed system is illustrated in Fig. 1, which outlines the step-by-step process for handling FER partial occlusion. After preprocessing, the images are partitioned into multiple regions (2, 3, 4, or 6), enabling the system to focus on localized areas that remain visible despite occlusion. This step is critical because certain facial regions, such as the eyes or upper face, often retain discriminative features even when the lower face is masked.

In the training phase, HoG features are extracted from each partitioned region to capture essential texture and edge information, which is then fed into SVM) classifiers. Each SVM is trained independently on its corresponding region, ensuring that localized characteristics are learned effectively. The independent models are then aggregated using an ensemble learning strategy with weighted soft voting, where weights are derived from the classification performance of each sub-model. This approach ensures that more reliable regions contribute more strongly to the final decision.

In the testing phase, the same preprocessing, partitioning, and feature extraction procedures are applied to unseen data. The outputs (probability vectors) from all regional SVM classifiers are aggregated through the ensemble mechanism to produce the final emotion prediction. This ensures consistency between training and inference, while also maintaining robustness against varying occlusion levels.

Finally, the performance of the proposed system is quantitatively evaluated using standard metrics, including accuracy, precision, recall, and F1-score. These metrics provide a comprehensive assessment of the system's effectiveness, capturing not only its correctness in emotion classification but also its ability to remain stable and reliable under different occlusion conditions. Together, these evaluations validate the robustness and generalization capability of the proposed PPM

combined with ensemble learning for FER under partial occlusion.

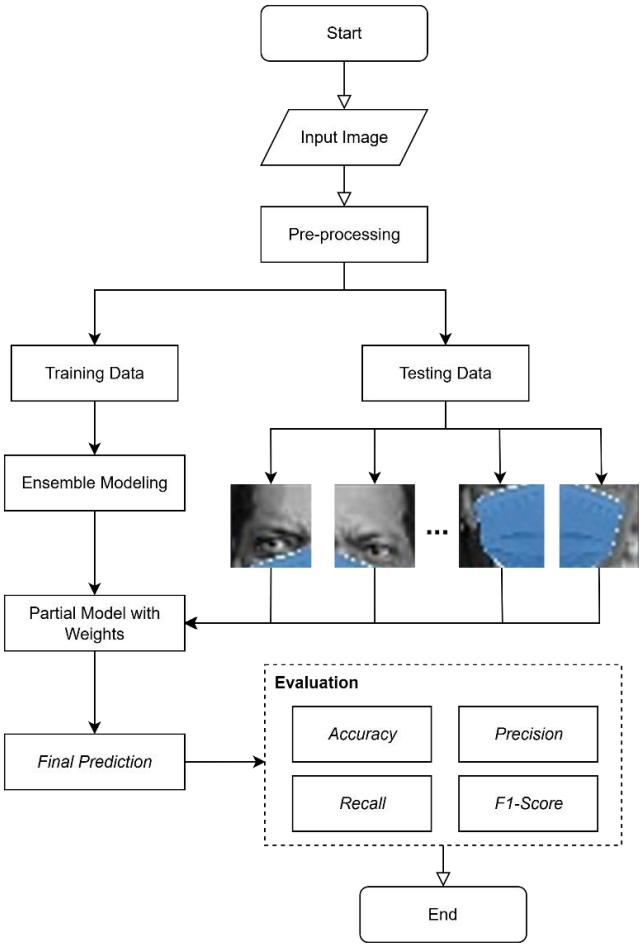


Fig. 1. Workflow of the algorithm.

B. Ensemble and Partial Part Models

Fig. 2 illustrates the architecture of the PPM using an Ensemble SVM approach for FER under partial occlusion [19]. The input facial image is first divided into multiple predefined regions (e.g., top-left, bottom-right), each capturing different parts of the face. These sub-regions are processed independently by separate SVM classifiers, with each model trained to recognize emotional patterns based on its corresponding part.

Each SVM generates a probability distribution over the emotion classes. These outputs are then aggregated using a soft voting strategy, where the final predicted label is determined by averaging the class probabilities across all SVMs and selecting the class with the highest average score. This strategy improves robustness by allowing the model to focus on visible and informative facial areas, compensating for parts affected by occlusion. It also ensures that even if certain regions are occluded or carry less expression-relevant information, other regions can effectively contribute to the final decision. The architecture is shown in Fig. 2.

C. Dataset and Preprocessing

The MaskedFER2013 dataset, containing 31,116 images sized 48×48 pixels, was used. Images were labeled into seven emotion classes: angry, disgust, fear, and many more. Each image underwent grayscale normalization and was partitioned into regions (2, 3, 4, or 6 parts) representing different facial areas [5]. This preprocessing phase is crucial, as it directly influences the quality of features that the model can extract. In this context, dominant features come from the eye and eyebrow regions, which typically display subtle variations across different emotions. Therefore, standardizing image size and format is essential to ensure consistent inputs for the classification model.

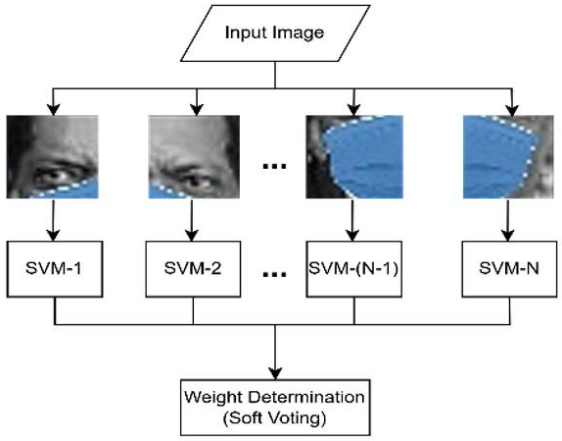


Fig. 2. Architecture of ensemble learning and partial part models.

Facial images were divided into several partitions to address the challenges posed by facial occlusion, particularly due to the use of masks. This partitioning strategy enables extracting localized features from different facial regions, which are combined to improve classification performance.

Fig. 3 illustrates the partitioning strategies applied to facial images with partial occlusion to improve emotion recognition. Each row represents an emotion class (Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise), while the columns show the original preprocessed image and its partitioned forms. The two-partition scheme divides the face into upper and lower regions, separating the eye and mouth areas. The three-partition scheme adds a middle section for finer detail. In contrast, the four-partition scheme splits the face into quadrants (top-left, top-right, bottom-left, bottom-right), enabling localized feature capture across both facial halves. This partitioning helps the model to differentiate features between the left and right sides of the face. The six-partition scheme further subdivides the image into top-left, top-middle, top-right, bottom-left, bottom-middle, and bottom-right regions, providing a highly localized representation at the cost of global context. These strategies enable complementary regional feature extraction, which enhances recognition performance under occlusion when integrated via ensemble learning. Overall, this figure demonstrates how different partitioning strategies will allow the model to capture both local and global discriminative features from unoccluded regions of the face. By leveraging these complementary partitions through ensemble learning, the system becomes more robust in recognizing emotions despite partial occlusion caused by masks. The results of the processed images and partition are shown in Fig. 3.

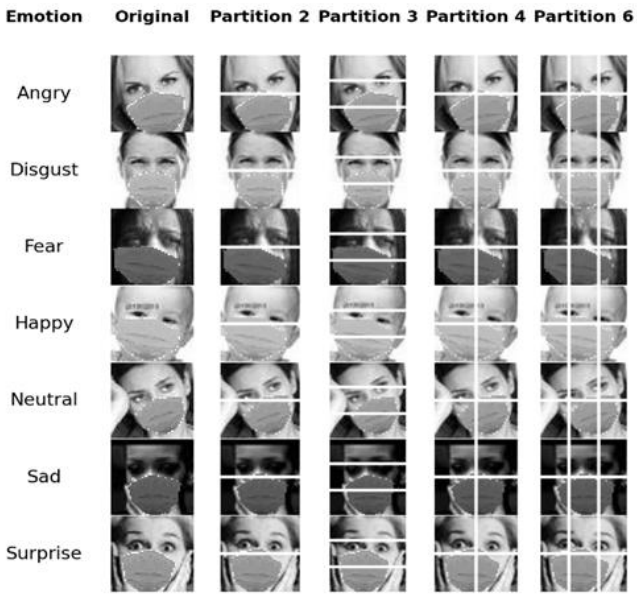


Fig. 3. Preprocessing and partition results.

D. Feature Extraction

In this stage, features were extracted from each partitioned image using the HoG method. HoG is particularly effective for capturing local edge orientations, gradient intensity, and texture patterns essential for distinguishing facial expressions, even under partial occlusion such as masks. By focusing on edge structures rather than pixel intensity, HoG can represent critical facial cues (e.g., eye contour, eyebrow shape, and forehead wrinkles) proper for emotion recognition. Each image partition produced a feature descriptor vector, which was later used as input for the classifiers [20].

E. Ensemble Prediction

Each SVM classifier generated prediction probabilities for the possible emotion classes during the testing phase. Instead of relying on a single classifier, the outputs from all SVMs were aggregated using a weighted soft voting mechanism. In soft voting, the predicted probabilities from individual classifiers were averaged, and the class with the highest average probability was selected as the final decision. This ensemble approach provided robustness, combining complementary information from different facial regions, reducing the impact of occlusion or local noise. Consequently, the ensemble prediction often yielded better overall performance compared to relying on a single model [21].

IV. Experimental Results

In this study, the dataset used is MaskedFER2013, which includes grayscale images with masks applied to simulate real-world occlusions. The dataset is split into two main groups to build and evaluate the model effectively: training data and testing data. From the full dataset, 80% is allocated to the training set and 20% to the testing set. Furthermore, the training set is subdivided into training and validation data, with an 80:20 ratio. The training data is used to train each SVM model on each facial partition. In contrast, the validation data is used during hyperparameter tuning to determine the optimal configuration for each SVM classifier before final training.

A. The Effect of Hyperparameter Tuning

A comprehensive hyperparameter tuning process was conducted to optimize the performance of SVM classifiers across different image partitioning strategies. The key hyperparameters tuned include the kernel type, regularization parameter (C), and kernel coefficient (gamma).

For the partial two-region approach, which divides the image into top and bottom sections, both models achieved the best results using the RBF kernel. The SVM Top model performed best with C = 5 and gamma = 1, reaching an accuracy of 31.19%, while the SVM Bottom model used C = 1 and gamma = 1, achieving 30.12%. This result indicates that both regions still hold significant emotional information despite mask occlusion, with the top region slightly outperforming the bottom. The optimal hyperparameter for the partial 2-Region SVM is shown in Table I.

TABLE I. Optimal Hyperparameter for Partial 2 SVM

Model	Kernel	C	Gamma	Accuracy
SVM Top	Linear	1	0.01	27.57%
	Poly	5	0.1	28.25%
	Rbf	**5**	**1**	**31.19%**
	Sigmoid	5	0.1	26.81%
SVM Bottom	Linear	1	0.1	26.99%
	Poly	1	0.1	28.09%
	Rbf	**1**	**1**	**30.12%**
	Sigmoid	0.1	0.01	26.51%

In the three-region partitioning (Top, Mid, and Bottom), the RBF kernel again yielded the best performance across all sections. The SVM Mid model stood out with C = 1 and gamma = 1, achieving an accuracy of 30.54%, higher than the Top (28.96%) and Bottom (28.13%) regions. This result implies that the middle portion of the face, typically including the eyes and nose bridge, contains richer emotional features under partial occlusion. The optimal hyperparameter for the partial 3-Region SVM is shown in Table II.

TABLE II. Optimal Hyperparameter for Partial 3 SVM

Model	Kernel	C	Gamma	Accuracy
SVM Top	Linear	0.1	0.01	26,47%
	Poly	5	0.1	27.19%
	Rbf	**5**	**1**	**28.96%**
	Sigmoid	0.1	0.01	26.51%
SVM Mid	Linear	1	0.01	27.57%
	Poly	5	0.1	29.52%
	Rbf	**1**	**1**	**30.54%**
	Sigmoid	0.1	0.01	26.51%
SVM Bottom	Linear	0.1	0.01	26.51%
	Poly	1	0.1	26.65%
	Rbf	**1**	**1**	**28.13%**
	Sigmoid	0.1	0.01	26.51%

The four-region partitioning (Top Left, Top Right, Bottom Left, and Bottom Right) showed a consistent trend where all best-performing models used the RBF kernel with C = 5 and gamma = 1. Accuracies in this configuration ranged from 29.64% to 31.05%, indicating symmetrical and localized facial features still provide valuable information for classification when the face is divided into quadrants. The optimal hyperparameters for the partial 4-Region SVM are shown in Table III.

TABLE III. OPTIMAL HYPERPARAMETER FOR PARTIAL 4 SVM

Model	Kernel	C	Gamma	Accuracy
SVM Top Left	Linear	0.1	0.01	28.13%
	Poly	1	0.1	30.14%
	Rbf	**5**	**1**	**30.72%**
	Sigmoid	1	0.1	26.63%
SVM Top Right	Linear	0.1	0.1	28.37%
	Poly	1	0.1	30.36%
	Rbf	**5**	**1**	**31.05%**
	Sigmoid	0.1	0.01	26.51%
SVM Bottom Left	Linear	1	0.1	27.67%
	Poly	1	0.1	28.68%
	Rbf	**5**	**1**	**29.66%**
	Sigmoid	0.1	0.01	26.51%
SVM Bottom Right	Linear	0.1	0.1	27.27%
	Poly	1	0.1	29.36%
	Rbf	**5**	**1**	**29.64%**
	Sigmoid	1	0.1	26.67%

The RBF kernel again dominated in performance in the most detailed six-region partitioning (Top Left, Top Mid, Top Right, Bottom Left, Bottom Mid, and Bottom Right). Most regions achieved optimal results with C = 5 and gamma = 1, except for the Top Mid region, which achieved better accuracy using C = 1. This suggests that the Top Mid region, including the eyes, eyebrows, and forehead, contains more complex features and benefits from lower regularization to avoid overfitting. Accuracies in this configuration ranged from 28.56% to 30.22%. The optimal hyperparameter for the partial 6-Region SVM is shown in Table IV.

Overall, the RBF kernel consistently produced the highest accuracies across all partitioning schemes, highlighting its effectiveness in modeling the non-linear nature of facial expression features. On the contrary, the sigmoid kernel showed the poorest performance in every scenario, indicating it is not suitable for this task. These findings support the hypothesis that local facial analysis, especially under occlusion like face masks, can reveal hidden expression patterns when combined with appropriate kernel functions and parameter settings.

B. Result of the Evaluation Model

To further evaluate the effectiveness of different approaches in handling FER under partial occlusion, a comparative analysis was conducted across several baseline CNN models, a non-partial SVM, and ensemble-based partial SVM. Table V summarizes the performance comparison of each model and partitioning strategy.

TABLE IV. OPTIMAL HYPERPARAMETER FOR PARTIAL 6 SVM

Model	Kernel	C	Gamma	Accuracy
SVM Top Left	Linear	1	1	27.15%
	Poly	1	0.1	29.02%
	Rbf	**5**	**1**	**30.22%**
	Sigmoid	0.1	0.01	26.51%
SVM Top Mid	Linear	0.1	0.1	28.31%
	Poly	**1**	**0.1**	**29.70%**
	Rbf	5	1	29.38%
	Sigmoid	1	0.1	26.99%
SVM Top Right	Linear	0.1	0.01	26.95%
	Poly	1	0.1	28.96%
	Rbf	**5**	**1**	**29.94%**
	Sigmoid	5	0.01	26.55%
SVM Bottom Left	Linear	0.1	0.1	26.81%
	Poly	1	0.1	27.37%
	Rbf	**5**	**1**	**28.66%**
	Sigmoid	5	0.01	26.53%
SVM Bottom Mid	Linear	0.1	0.1	28.11%
	Poly	1	0.1	29.10%
	Rbf	**1**	**1**	**30.08%**
	Sigmoid	5	0.01	27.83%
SVM Bottom Right	Linear	0.1	0.1	26.65%
	Poly	1	0.1	27.41%
	Rbf	**5**	**1**	**28.56%**
	Sigmoid	1	0.1	26.53%

TABLE V. RESULT COMPARISON EVALUATION MODEL

Experiment	Accuracy	Precision	Recall	F1-Score
Non-Partial VGG16	25%	15%	15%	10%
Non-Partial ResNet50	14%	2%	14%	3%
Non-Partial MobileNetV2	21%	14%	15%	13%
Non-Partial SVM	44%	50%	40%	45%
Partial 2 SVM	43%	55%	37%	40%
Partial 3 SVM	40%	59%	33%	34%
Partial 4 SVM	**45%**	61%	39%	42%
Partial 6 SVM	43%	67%	36%	38%

Table V presents the comparative evaluation results of several non-partial deep learning models (VGG16, ResNet50, and MobileNetV2), a non-partial SVM model, and multiple ensemble partial SVM models. These findings suggest that without partial strategies or SVM integration, CNN models

alone are less effective in recognizing facial emotions under occlusion conditions such as masks.

In contrast, the SVM approach delivers significantly more stable results. The Non-Partial SVM model achieves 44% accuracy and 45% F1-score, nearly double the performance of CNN-based models. The relatively higher F1-score indicates that the SVM maintained a better balance between precision and recall, even though the dataset involved masked faces. This finding can be attributed to SVM's ability to work well with small- to medium-sized datasets and handcrafted features (e.g., HoG), which are less data-hungry compared to CNNs. Using non-partitioned images allowed the SVM to capture global facial information (both visible and partially occluded regions), providing a more holistic representation for classification.

Furthermore, ensemble-based partial SVM models reveal distinct performance characteristics. The Partial 2 SVM records 43% accuracy with 55% precision, effectively reducing false positives. Partial 3 SVM shows lower recall (33%) but higher precision (59%), although its F1-score remains relatively modest at 34%. Partial 4 SVM stands out with the highest accuracy (45%) and balanced performance across metrics, achieving 61% precision, 39% recall, and 42% F1-score. Meanwhile, Partial 6 SVM achieves the highest precision (67%), but with lower recall (36%), resulting in an F1-score of 38%.

The F1-score is the harmonic mean of precision and recall, reflecting the balance between these metrics. The Non-Partial SVM achieved the highest F1-score (45%) because it had access to the complete facial representation, enabling it to capture both global and local cues, even when some regions were occluded. In contrast, partition-based models suffered from an imbalance: while precision improved (e.g., 67% for the 6-partition SVM), recall decreased significantly (e.g., 36%), leading to a lower F1 score. This result indicates that partition-based models were more "conservative" in their predictions (fewer false positives), but they missed many true cases (higher false negatives). The holistic non-partial approach helped the SVM achieve a better trade-off, resulting in the highest F1-score overall.

C. The Effect of Ensemble

Ensemble learning was employed in this study to enhance the robustness of FER under partial occlusion. Instead of relying on a single global model, multiple SVM classifiers were trained independently on different facial partitions. Each partition captures localized features from specific regions of the face (e.g., eyes, forehead, or partial cheek areas), which are particularly relevant when masks occlude other regions. During inference, the outputs of these individual classifiers were aggregated using the weighted soft voting technique.

A generalized weighted soft voting scheme was adopted to ensure that the ensemble method remains applicable across different partitioning strategies (two, three, four, and six regions). Let each partition model $i \in \{1, \dots, N\}$ produce a class probability distribution $P_i(c)$ for class $c \in C$. The relative contribution of each model is determined by its validation performance score s_i, which represents accuracy. These scores are normalized into weights as follows:

$$\omega = \frac{s_i}{\sum_{j=1}^{N} S_j}, \quad i = 1, \dots, N \tag{1}$$

The ensemble probability distribution is then obtained through the weighted summation of the outputs of all partition models:

$$P_{ensemble}(c) = \sum_{i=1}^{N} \omega_i P_i(c), \quad \forall c \in C \tag{2}$$

Finally, the predicted class label is assigned based on the maximum ensemble probability:

$$\hat{y} = \arg\max_{c \in C} P_{ensemble}(c) \tag{3}$$

This formulation enables the ensemble strategy to flexibly adapt to different partitioning schemes without modification to the core framework. In the case of equal weighting, setting all $s_i = 1$ yields the traditional unweighted soft voting. By leveraging performance-driven weighting, partitions that demonstrate higher discriminative capability are assigned greater influence in the final decision, thereby improving robustness under partial occlusion scenarios.

Table VI to Table IX summarize the comparative performance between the single partial models and the ensemble approach. The results indicate that the ensemble strategy provides higher overall accuracy and better macro-level metrics compared to individual models. This result shows that weighted soft voting effectively combines complementary strengths from different facial regions. Consequently, the ensemble approach achieves a more balanced and reliable performance across all emotion classes.

TABLE VI. RESULT COMPARISON PARTIAL 2 MODEL

Model	Accuracy	Precision	Recall	F1-Score
SVM Top	0.42	0.50	0.38	0.40
SVM Bottom	0.37	0.44	0.32	0.33
Ensemble Partial 2	0.43	0.55	0.37	0.40

TABLE VII. RESULT COMPARISON PARTIAL 3 MODEL

Model	Accuracy	Precision	Recall	F1-Score
SVM Top	0.38	0.48	0.33	0.35
SVM Mid	0.38	0.45	0.32	0.34
SVM Bottom	0.35	0.48	0.28	0.29
Ensemble Partial 3	**0.40**	**0.59**	**0.33**	**0.34**

TABLE VIII. RESULT COMPARISON PARTIAL 4 MODEL

Model	Accuracy	Precision	Recall	F1-Score
SVM Top Left	0.44	0.52	0.39	0.42
SVM Top Right	0.43	0.51	0.38	0.41
SVM Bottom Left	0.36	0.45	0.30	0.31
SVM Bottom Right	0.37	0.44	0.30	0.31
Ensemble Partial 4	**0.45**	**0.61**	**0.39**	**0.42**

TABLE IX. RESULT COMPARISON PARTIAL 6 MODEL

Model	Accuracy	Precision	Recall	F1-Score
SVM Top Left	0.41	0.52	0.37	0.40
SVM Top Mid	0.36	0.41	0.30	0.31
SVM Top Right	0.41	0.52	0.37	0.39
SVM Bottom Left	0.36	0.47	0.29	0.30
SVM Bottom Mid	0.35	0.46	0.28	0.28
SVM Bottom Right	0.37	0.48	0.30	0.31
Ensemble Partial 6	**0.43**	**0.67**	**0.36**	**0.38**

The evaluation of different partitioning strategies (Table VI to Table IX) reveals varying levels of effectiveness in handling partial occlusion. The two-partition ensemble achieves only moderate performance (accuracy = 0.43), as the division into upper and lower regions often limits feature diversity when masks heavily occlude the lower half. Similarly, the three-partition ensemble slightly improves accuracy to 0.40. However, the finer segmentation does not yield significant performance gains, as critical features may still be missing in one or more regions.

The six-partition scheme demonstrates competitive precision (0.67); however, its overall accuracy (0.43) and F1 score (0.38) remain lower, likely due to the fragmentation of global facial context into overly localized regions. In contrast, the four-partition ensemble strikes the most effective balance by capturing vertical and horizontal facial feature variations. With an accuracy of 0.45 and macro precision of 0.61, it consistently outperforms other partition strategies, confirming that the quadrant-based division provides complementary information while preserving sufficient contextual cues.

Therefore, the four-partition ensemble is identified as the most robust approach for FER under partial occlusion, combining enhanced accuracy with improved precision while maintaining stable recall and F1-score performance.

D. Discussion

This study conducted a series of facial expression classification experiments using the SVM algorithm with different approaches: non-partial, partial two-part, partial three-part, partial four-part, and partial six-part ensembles. The hyperparameter tuning process consistently showed that the best-performing configuration was C=5, gamma=1, with the RBF kernel, regardless of the number of partitions. For example, in the four-part approach (Top Left, Top Right, Bottom Left, Bottom Right), all models converged to the same parameter set.

From the evaluation results, the Non-Partial SVM achieved the highest F1-score (43%), reflecting a balanced trade-off between precision and recall. This result indicates that the model is accurate and sensitive to detecting a wide range of classes, making it suitable for general-purpose applications such as human-computer interaction or assistive technologies. Such performance aligns with the findings of [12], who also reported that traditional CNN and SVM models perform well under full-face visibility but degrade substantially when occlusion increases.

In contrast, the Partial 6 SVM demonstrated the highest precision (67%), making it more reliable for applications requiring high-confidence predictions with minimal false positives. This conservative approach, although sacrificing recall, is valuable in contexts like early emotion disorder detection or security systems. Meanwhile, the Partial 4 SVM achieved the highest accuracy (45%), producing the most significant number of correct classifications overall. Although its F1-score is slightly lower than the non-partial model, this strength makes it suitable for scenarios where maximizing the number of accurate predictions is the main priority. Similar behavior was reported by [13], where models prioritizing confident regions yielded higher precision but lower recall under occluded conditions.

Meanwhile, the Partial 4 SVM obtained the highest accuracy (45%), demonstrating a balanced trade-off between model complexity and region diversity. This configuration benefits from adequate regional representation (Top Left, Top Right, Bottom Left, Bottom Right) without introducing excessive fragmentation that might dilute local feature learning. The performance consistency across partitions suggests that the ensemble learning mechanism, through weighted soft voting, effectively integrates localized cues from multiple facial regions, thus maintaining reliability even under partial occlusion. This finding supports the concept of region-based feature fusion as highlighted in [11], which emphasized the importance of focusing on visible regions to improve recognition accuracy.

To ensure that the model's performance is consistent and not biased toward a specific subset of the data, a 3-fold cross-validation was applied automatically during the hyperparameter tuning process using GridSearchCV from scikit-learn. This mechanism divides the dataset into three parts, where two folds are used for training and one for validation in each iteration. The cross-validation process helps evaluate the model's generalization ability while selecting the best parameters for the Support Vector Machine (SVM).

To further analyze class-wise performance and model consistency, confusion matrices were generated for each regional SVM classifier, as shown in Fig. 4 to Fig. 7. The matrices illustrate that the model performs consistently across multiple emotion categories, with the highest correct classifications observed in neutral and happy classes. This analysis confirms that the model maintains balanced recognition performance across visible facial regions despite partial occlusion.

These matrices clearly show that the "happy" class dominates the correctly classified samples, which aligns with the fact that the "happy" expression has the largest number of samples in the MaskedFER2013 dataset. This class imbalance makes the model more confident in recognizing "happy" features, even under partial occlusion. In contrast, the "disgust" and "fear" classes exhibit noticeably lower recognition rates, often being misclassified as sad or angry. This misclassification trend occurs because these emotions share similar upper-face features (such as eyebrow contraction), which become more dominant when the lower face is covered by a mask.

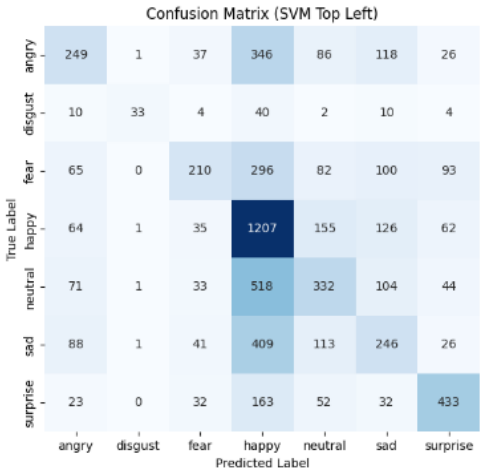


Fig. 4. Confusion matrix (Top Left Partial 4 SVM).

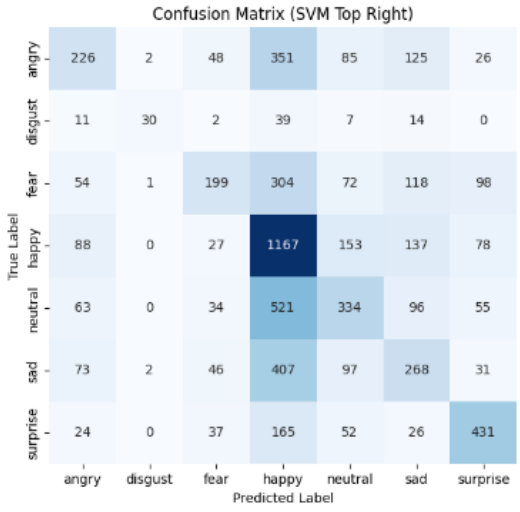


Fig. 5. Confusion matrix (Top Right Partial 4 SVM).

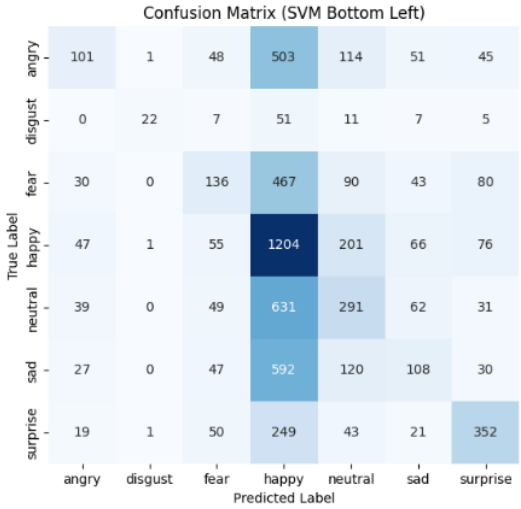


Fig. 6. Confusion matrix (Bottom Left Partial 4 SVM).

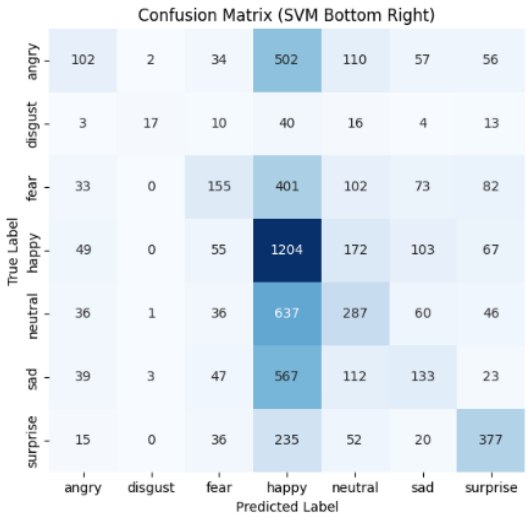


Fig. 7. Confusion matrix (Bottom Right Partial 4 SVM).

Additionally, the "neutral" and "surprise" expressions display moderate accuracy, as these emotions retain distinctive cues around the eye and forehead regions that remain visible even with occlusion. Misclassifications between neutral and happy are also frequent, indicating that when the mouth region is hidden, the model sometimes relies too heavily on the upper-face features, leading to overlapping interpretations.

The relatively consistent diagonal dominance across the four regional SVM (Top Left, Top Right, Bottom Left, and Bottom Right) confirms that the approach maintains robust and regionally stable behavior. This suggests that the ensemble structure successfully mitigates the performance degradation typically caused by partial occlusion, allowing the system to retain a reasonable level of accuracy and interpretability.

Despite these promising results, several limitations remain. First, the proposed method relies heavily on handcrafted features (HOG), which may not fully capture complex spatial and emotional variations compared to deep learning representations. Second, the partitioning strategy is fixed and non-adaptive; dynamic or attention-based segmentation could better capture relevant visible regions under real-world occlusion patterns. Finally, the dataset used (MaskedFER2013) represents a constrained environment with limited occlusion diversity, which may not generalize to unconstrained or in the wild scenarios.

For future research should focus on integrating deep feature extraction with the ensemble SVM architecture, enabling richer and more abstract representations of partially visible faces. Additionally, adaptive region partitioning using attention maps or saliency detection could further enhance robustness by dynamically identifying the most informative facial regions. Expanding the evaluation to include cross-dataset testing (e.g., MaskedFER2023 or MaskedCK+) will also help assess generalization capability. Lastly, incorporating temporal information from video sequences could provide a more comprehensive understanding of emotional dynamics under occlusion.

V. Conclusion

Based on the comparative analysis, no single model can be regarded as the absolute best across all evaluation metrics. Instead, each ensemble configuration demonstrates specific contextual strengths depending on the performance metric prioritized in real-world applications. The Non-Partial SVM model shows a more balanced performance regarding F1-score and recall, making it well-suited for general-purpose use, where both correctness and coverage are equally important. In contrast, the Partial-6 SVM ensemble achieves the highest precision, which indicates its suitability for applications that demand high-confidence decision-making, where minimizing false positives is a priority. Meanwhile, the Partial-4 SVM ensemble yields the highest overall accuracy, suggesting that it is most effective in scenarios where maximizing the total number of correct classifications is the primary objective. These findings highlight that model selection should not rely solely on a single metric, but rather be aligned with the intended application context and performance requirements.

Acknowledgment

The research and publication were funded through the Final Assignment Recognition Program (Program Rekognisi Tugas Akhir) Batch II Year 2022, Universitas Gadjah Mada with Assignment Number 633/UN1.P.III/KPT/HUKOR/2022 and 5722/UN1.P.III/Dit-Lit/PT.01.05/2022.

References

[1] Y. Li, J. Wei, Y. Liu, J. Kauttonen, and G. Zhao, "Deep Learning for Micro-Expression Recognition: A Survey," IEEE Trans. Affect. Comput., vol. 13, no. 4, pp. 2028–2046, 2022, doi: 10.1109/TAFFC.2022.3205170.

[2] V. Bettadapura, "Face Expression Recognition and Analysis: The State of the Art," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 12, pp. 1424–1445, 2012, doi: 10.1109/34.895976.

[3] C. Ramdani, M. Ogier, and A. Coutrot, "Communicating and reading emotion with masked faces in the Covid era : A short review of the literature," Psychiatry Res., vol. 316, no. January, p. 114755, 2022, doi: 10.1016/j.psychres.2022.114755.

[4] D. B. Grahlow M, Rupp CI, "The impact of face masks on emotion recognition performance and perception of threat," pp. 1–16, 2022, doi: 10.1371/journal.pone.0262840.

[5] C. Jiang, R. Hasan, T. Gedeon, and Z. Hossain, "MaskTheFER : Mask-Aware Facial Expression Recognition using Convolutional Neural Network," 2023 Int. Conf. Digit. Image Comput. Tech. Appl., pp. 456–463, 2023, doi: 10.1109/DICTA60407.2023.00069.

[6] M. P. K. Putra and Wahyono, "A Novel Method for Handling Partial Occlusion on Person Re-identification using Partial Siamese Network," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 7, pp. 313–321, 2021, doi: 10.14569/IJACSA.2021.0120735.

[7] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," Proc. - Int. Conf. Image Process. ICIP, vol. 2, no. March 2015, pp. 370–373, 2005, doi: 10.1109/ICIP.2005.1530069.

[8] G. Priyanka and S. Pavithra, "Facial Expression Recognition using SVM With CNN and Handcrafted Features," Int. J. Recent Technol. Eng., vol. 8, no. 4, pp. 3570–3574, 2019, doi: 10.35940/ijrte.d7802.118419.

[9] S. Zhang, X. Zhao, and B. Lei, "Robust facial expression recognition via compressive sensing," Sensors, vol. 12, no. 3, pp. 3747–3761, 2012, doi: 10.3390/s120303747.

[10] S. R. Thavarekere, A. Hebbar, and D. Uma, "A Deep Learning Approach to Facial Expression Recognition in the Presence of Masked Occlusion," INDICON 2022 - 2022 IEEE 19th India Counc. Int. Conf., pp. 1–7, 2022, doi: 10.1109/INDICON56171.2022.10040209.

[11] Y. Wang, Kai and Peng, Xiaojiang and Yang, Jianfei and Meng, Debin and Qiao, "Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition," Trans. Img. Proc., vol. 29, pp. 4057–4069, 2020, doi: https://doi.org/10.1109/TIP.2019.2956143.

[12] Y. Li, S. Member, J. Zeng, and S. Shan, "Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism," IEEE Trans. Image Process., vol. 28, no. 5, pp. 2439–2450, 2019, doi: 10.1109/TIP.2018.2886767.

[13] H. Ding, P. Zhou, and R. Chellappa, "Occlusion-Adaptive Deep Network for Robust Facial Expression Recognition," IEEE Int. Jt. Conf. Biometrics, 2020, doi: 10.1109/IJCB48548.2020.9304923.

[14] N. Y. Abdullah and A. M. F. Alkababji, "Masked face with facial expression recognition based on deep learning," Indones. J. Electr. Eng. Comput. Sci., vol. 27, no. 1, pp. 149–155, 2022, doi: 10.11591/ijeecs.v27.i1.pp149-155.

[15] I. J. H. Recto, "Synthetic Occluded Masked Face Recognition using Convolutional Neural Networks," 2022 IEEE Int. Conf. Ind. 4.0, Artif. Intell. Commun. Technol., pp. 124–129, 2022, doi: 10.1109/IAICT55358.2022.9887517.

[16] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," 2010 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Work. CVPRW 2010, no. July, pp. 94–101, 2010, doi: 10.1109/CVPRW.2010.5543262.

[17] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with Gabor wavelets," Proc. - 3rd IEEE Int. Conf. Autom. Face Gesture Recognition, FG 1998, pp. 200–205, 1998, doi: 10.1109/AFGR.1998.670949.

[18] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," Neural Networks, vol. 64, pp. 59–63, 2015, doi: 10.1016/j.neunet.2014.09.005.

[19] L. Breiman, "Bagging predictors," Risks, vol. 24, no. 3, pp. 123–140, 1996, doi: https://doi.org/10.1007/BF00058655.

[20] R. Noviyanti and E. Sinduningrum, "Braille Character Recognition with Histogram of Oriented Gradients (HOG) and SVM-Based Image Processing," Syntax Lit. ; J. Ilm. Indones., vol. 9, no. 8, pp. 4411–4419, 2024, doi: 10.36418/syntax-literate.v9i8.16951.

[21] A. Manconi, G. Armano, M. Gnocchi, and L. Milanesi, "A Soft-Voting Ensemble Classifier for Detecting Patients Affected by COVID-19," Appl. Sci., vol. 12, no. 15, 2022, doi: 10.3390/app12157554.

Ambulance Detection and Priority Passage at Urban Intersections Using Transfer Learning and Explainable AI

Murtaza Hanif[1], Taj Muhammad[2], Atif Ikram[3,*], Shahid Yousaf[4], Marwan Abu-Zanona[5], Asef Mohammad Ali Al Khateeb[6], Bassam Elzaghmouri[7], Saad Mamoun Abdel Rahman Ahmed[8], Lamia Hassan Rahamatalla[9]

Department of Computer Engineering, University of Engineering and Technology, UET, Lahore, 54890, Punjab, Pakistan[1, 2]
Department of Computer Science & IT, University of Lahore, 54890, Punjab, Pakistan[3, 4]
College of Information Technology, Amman Arab University, Amman, 11118, Jordan[3]
Department of Management Information systems-College of Business Administration, King Faisal University, Al-ahsa, Saudi Arabia[5, 9]
Department of Computer Science, Faculty of Computer Science and Information Technology, Jerash, Jordan[6]
Department of Data Science and Artificial Intelligence, Al-Ahliyya Amman University, Amman 19328[7]
Applied College, King Faisal University, Al-Ahsa, Saudi Arabia[8]

***Abstract*—Static traffic signal timings often cause severe delays for emergency vehicles, including ambulances at junctions in urban areas, putting lives at risk. To highlight this, the present study proposes an intelligent traffic control system that dynamically adjusts traffic signals based on real-time monitoring. The system employs a yolov8-based deep learning model fine-tuned through transfer learning for ambulance detection from live video. At an Intersection over Union (IoU) threshold of 0.5, the model achieves a mean Average Precision (mAP) of 0.860. To ensure continuous tracking, NORFair tracking is implemented to ensure constant detection across frames. Additionally, to improve explainability and, the frame incorporates Local Interpretable Model-Agnostic Explanation (LIME), providing visual signals into the model decision-making process. Once an ambulance is detected, the system instantly triggers a green-light activation for the ambulance's lane, enabling quick emergency response. Unlike conventional systems with fixed signal timing, this approach enables smart and adaptive traffic management in urban environment. However, the system's shortcomings in low-visibility situations, such as at night or in fog, despite its encouraging results, highlight the need for incorporating images taken at night and in foggy weather into the dataset.**

Keywords*—Ambulance detection; YOLOV8; LIME; transfer learning; NorFair; urban area; traffic control; smart traffic management

I. Introduction

An efficient emergency response is crucial for saving lives, especially in developing countries where road networks struggle to keep pace with the increasing number of vehicles. Emergency response, which includes ambulances [1] and bikes [2], often faces significant delays due to heavy traffic congestion, particularly at intersections. Traditional traffic management systems operate on pre-installed cycles, failing to dynamically adapt to real-time emergencies, which minimizes response times and hinders the traversal of ambulances through intersections. As the number of vehicles increases by a million every decade [3], it has become a concern for many researchers to provide quality traffic management, particularly for ambulances, to mitigate response times for emergencies.

Consequently, Machine learning and computer vision have unexpectedly enhanced intelligent systems, such as object detection [4], object tracking [5], vehicle counting [6], and traffic light detection [7]. Similar to detection and identification, computer vision algorithms have been adopted to calculate the vehicle speed [8] and pollution monitoring [9]. Furthermore, Machine learning algorithms are employed to control the path following of unmanned vehicles [10], robot navigation [11], and adaptive signal control [12]. While Deep learning models such as YOLO (You Only Look Once) [13] and R-CNN (Region-Based Convolutional Neural Networks) [14] have been effectively utilized for object detection, including the identification of emergency vehicles [15], adapting audio based location of vehicle [16]. Existing research primarily focuses on detection rather than integrating this detection into real-time, dynamic traffic control decision. Most prior studies lack the ability to autonomously manipulate traffic signals in response to ambulance localization, resulting in delayed emergency response and inefficient intersection management. Furthermore, explainable AI techniques have rarely been employed to interpret or validate the model's decision-making process in such safety-critical applications

The primary study of this article is based on an AI-driven traffic signal optimization framework that dynamically modifies vehicle routes upon detecting an approaching ambulance. The method adopted involves leveraging the LIME technique to enhance the confidentiality of the localization system while incorporating the tracking technique to improve traversal through the intersection. The core of this research lies in adopting a tailored object localization model, Yolov8, as a transfer learning approach. The model demonstrates outstanding performance, achieving a mean average precision of 0.85 at the threshold of 0.5. To improve signal optimization

beyond vehicle detection, tracking helps monitor the vehicle's trajectory and movement over time. Fig. 1 illustrates the overall system flow.

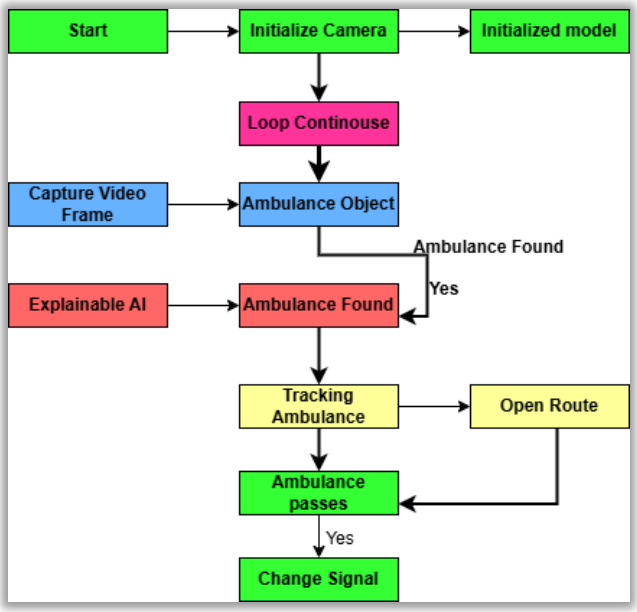


Fig. 1. The flow of the system.

The article presents a novel approach to improving ambulance localization explainability for optimizing ambulance traversal. The contribution of this article can be summarized as follows.

1) Model modification and adaptation: A customized Yolov8 model was used for ambulance detection in a dynamic and active environment using transfer learning, offering robust accuracy and efficiency. Formerly, the model was customized for object localization to support navigation for visually impaired people. This deliberate modification is intended to reduce the parameter, making the model effective for operation, even for a small dataset.

2) LIME for explainability: A significant novelty of this work is the amalgamation of LIME (Local Interpretable Model-Agnostic Explanation) [17] to provide understanding into the model's decision-making process. Integrating LIME into the localization model's prediction, we were able to understand and explain the spatial regions of the image that meaningfully influenced the model's decision to categorize an object as an ambulance.

3) Model transparency: We discussed the black-box nature of the deep learning model by applying LIME [17], making the localization model's predictions more understandable. The LIME explanations discovered which specific visual features helped the model's detection of an ambulance, offering transparency into the decision-making process.

4) Validation and real-world scenario: We demonstrate the practical utility of this approach by using the model on real-world video, including complex situations where ambulances are shown partially. This demonstrates the capability of the detection model, combined with LIME, to provide both accurate and explainable results across various conditions, making the approach suitable for deployment in real-world applications.

5) Optimizing the ambulance traversal: Unlike the existing system, which depends on fixed turns. Our approach focuses on enabling autonomous signal manipulation based on emergency or ambulance detection.

Section II of this article presents a literature review, providing a comprehensive examination of prior literature. Section III details the materials & methods, including dataset justification and selection, as well as the adopted methodology and training process. Additionally, Section IV presents the performance evaluation of the proposed framework, outlining the metrics and parameters used to assess its effectiveness. Section V highlights the results derived from the implemented methodology, while Section VI offers a detailed discussion and interpretation of these findings in relation to existing studies. Finally, Section VII concludes the paper with insights and outlines potential future direction for improving system performance under challenging conditions such as low visibility or heavy traffic.

II. Literature Survey

The escalation of traffic congestion in urban areas, particularly in developing countries, has brought significant challenges to the emergency response system [18]. Traditional traffic signal system [19], operates on predefined fixed-time cycles, often failing to adapt dynamically to real-time scenarios, which leads to delays in emergency vehicle transit such as ambulance. While these conventional systems are simple and cost effective, their lack of adaptability makes them unsuitable for modern, high density urban networks where emergency response time is critical.

Previous research highlights the successful application of artificial intelligence and its various branches across diverse fields [20], [21], [22], [23], [24], [25]. In the context of intelligent transport system (ITS) deep learning models have been employed for task such as object localization, classification and traffic pattern prediction,[4], [26]. Among these YOLO (You Only Look Once) [13] and its successive versions have been recognized for their high speed and accuracy in real time object detection. The newest iteration, YOLOv8 [27], suggests enhanced precision and speed, making it suitable for reliable real-time applications in dynamic environments. Likewise, the model is used successfully and precisely for locating emergency vehicles [18], most studies stop at detection and don't address how this information can be used to actively manage to manipulate traffic signals in response. This limited their practical impact on improving emergency response time.

To refine ambulance detection, researchers have explored multimodal fusion approach that combines visual data with audio cues [16], such as a siren, enhancing detection accuracy under diverse environmental conditions. Furthermore, a hybrid system utilizing an LSTM network for audio signal processing and a ResNet-18 for visual data was suggested [27] have demonstrated higher localization accuracy under noisy or low

visibility conditions. These models, however, are often computationally expensive and difficult to deploy in real-time citywide traffic systems, especially in developing regions, where hardware resources are limited.

In addition to advancements in detection and tracking, the nature of deep learning models still raises concerns about the decision-making processes they employ which limit trust and interpretability in safety critical applications such as emergency response. To mitigate this, Explainable Artificial Intelligence (XAI) techniques, such as LIME [28], Grad-CAM [29], have been proposed to make model decision more interpretable. Recent studies have applied XAI in traffic signal optimization [30]. Allowing engineers to better understand and validate model decisions. However, most of these approaches remain experimental and have not yet been integrated into end-to-end real-time ambulance detection.

Results from integrating XAI approaches into traffic management systems have been encouraging [31]. Existing research still lacks a unified framework that combines all the real-time ambulance localization, tracking and interpretable model decisions to dynamically adjust traffic signal. Few studies have attempted to bridge detection and action translating ambulance recognition directly into automated signal manipulation. Addressing this gap, the present study our process focuses on leveraging costumed yolov8 for ambulance localization through transfer learning, incorporating a tracking algorithm for trajectory monitoring, while employing LIME for model Interpretability. This contributes to enhancing emergency response times by ensuring that the traffic signal dynamically adapts to the presence of an ambulance, facilitating their passage through the intersection.

III. Material and Methods

This research article presents an ambulance traversal mechanism through the intersection. The system enhances the dynamic manipulation of the signal by detecting vehicles. Additionally, it provides tracking for monitoring the ambulance's trajectory and LIME for a detailed understanding of why the vehicle was identified as an ambulance.

A. Dataset Overview

To train and evaluate our ambulance traversal system through the intersection, we utilized a publicly available dataset sourced from the Roboflow platform [https://universe.roboflow.com/himank-vpetc/ambulance-4bova/dataset/1]. This dataset was specifically selected for its relevance to real-world situations where the timely localization of vehicles, such as ambulances, is vital for dynamic signal manipulation and the quick passage of ambulances through intersections. The dataset comprises a diverse set of images featuring ambulances in various backgrounds, lighting conditions, and angles, providing robust variability that supports generalization during model training. In addition, the photos are annotated using rectangle bounding boxes that label instances of the ambulances, enabling our customized object detection model to learn spatial and visual features for efficient identification.

B. Dataset Specifications

1) Source: Roboflow Universe [31]

2) Category: Object detection (Single class "Ambulance")

3) Number of images: We used a dataset of 6,432 images, each with dimensions 640 × 640 pixels and 3 color channels, for training the object detection model.

4) Annotation Format: YOLO rectangle bounding box

5) Image Resolution: The images were down sampled and standardized to a resolution of 640 × 640 pixels

6) Environment Type: Urban outdoor scenes, roads

7) Data Split: 70% training, 20% validation, 10% test. The approach is employed to gather a balance trade-off between the training and validation performance evaluation and the final test accuracy.

8) Additionally, this uniform split ensures a consistent distribution across subsets, which aids in maintaining a representative sample and preventing overfitting.

C. Justification for Database Selection

The needs of our study, which centered on the real-time detection of ambulances at urban intersections to enable dynamic traffic signal control, led to the selection of the "Ambulance" dataset from Roboflow [31]. This dataset, specifically curated for ambulance recognition, differs from general-purpose object detection datasets in that it offers a focused and domain-relevant collection of annotated images under various real-world scenarios.

The dataset is ideal for training a robust detection model that can operate reliably in live surveillance setups, as ambulances are present in a variety of backgrounds, including junctions, roadside settings, and diverse lighting conditions. Since our suggested solution directly initiates the dynamic reconfiguration of traffic signals to prioritize the passage of emergency vehicles, accurate and timely ambulance recognition is essential.

This dataset aligns well with our research aim, operational, and technical objectives due to its excellent bounding box annotations and conformity to the customized ambulance detection training format. As a result, its applicability and compatibility with real-time, vision-based smart traffic control systems are.

D. Model Architecture and Training Process

1) Model selection and justification: In this article, we employed customized YOLOv8 architecture previously modified to highlight the challenges associated with limited data availability. To address the constraints posed by small dataset sizes, a deliberate customization was introduced by reducing the convolutional kernel size from the conventional 3×3 to a more compact 2×2 configuration. This systematic manipulation allows the model to analyze minute details more effectively while concurrently reducing the number of learnable parameters, thus mitigating the risk of overfitting and improving generalization performance.

Furthermore, the padding function was modernized to match the reduced kernel size, confirming proper padding and spatial consistency across layers. The customized convolutional layers combined 64 filters with a stride of 2 × 2 and padding of 1, keeping the dimension of the feature maps. Furthermore, the rectified linear unit (ReLU) activation function was used to increase the discriminative power of feature extraction.

This precisely customized layer design aimed to optimize feature representation while preserving model efficiency. Building upon this foundation, the customized yolov8 model was used through transfer learning to fine-tune it on the ambulance detection job.

E. Training Setup

For transfer learning, the model was retrained using an explicit configuration on the openly available ambulance detection dataset to facilitate transfer learning. The input frames were down sampled to 640 × 640 pixels, and the training process was carried out for 50 epochs, applying a CUDA GPU to quicken the process. To conserve formerly learned generic features, the first ten layers of the network were frozen, allowing only the subsequent layers to fine-tune for ambulance-specific localization. To minimize overfitting, Stochastic Gradient Descent (SGD) was employed, utilizing a learning rate of 0.01, a momentum of 0.9, and a weight decay of 0.001. In addition to visual inspection of predicted bounding boxes, the mean Average Precision (mAP) measure on the validation set was used to track model performance during training. Using an Intersection over Union (IoU) criterion of 0.5, the test set's final evaluation yielded encouraging results with a mean Average Precision (mAP) score of 0.85 shown in PR curve, indicating a high intersection ambulance detection ability.

During training, the model's performance on the validation set was continuously monitored using the mean average precision (mAP) metric. In addition, visual inspection of the model's predictions was conducted to qualitatively assess its accuracy. Upon completion of training, the test set was used to evaluate the model's performance using the same metrics as the validation set. Evaluation was performed at an intersection over union (IoU) threshold of 0.5. The trained model achieved an impressive mAP of 0.850, demonstrating its strong object detection capability. The results of the model are shown in Fig. 2.

Fig. 2. Test set.

For the evaluation of our intelligent traffic management system while incorporating real-time ambulance localization, tracking, with explainable AI, for efficient vehicle traversal, the camera continuously captures the video data to locate the ambulances from the intersection point. Capitalizing on the detection model and the Norfair tracker for persistent object localization—even during occlusions or temporary detection losses—the system reliably monitors ambulance movement. Upon detection, the system triggers the simulated traffic signal to transition to green, allowing for smooth passage. In addition to improving transparency, Local Interpretable Model-agnostic Explanations (LIME) produces visual rationales that capture the most influential frame region, which contributes to model classification. Periodic LIME explanations for high-confidence localization depict insight into the decision-making process, while the Norfair tracker maintains robust trajectory estimation by incorporating location across frames. This integration of state-of-the-art localization, persistent tracking, and explainable AI establishes a responsive and trustworthy emergency vehicle priority system for the urban environment shown in Real-time detection of an ambulance (Fig. 3).

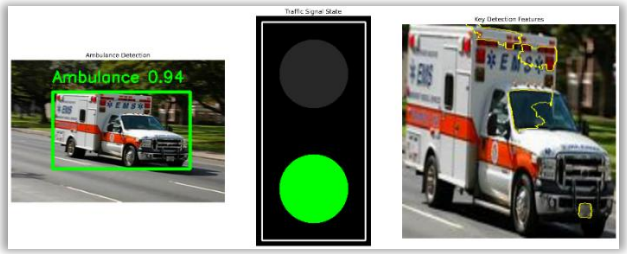


Fig. 3. Real-time detection of an ambulance.

This ambulance passage, combined with ambulance localization, tracker, and LIME, aims to enhance the dynamic management of traffic lights, allowing for reducing response time and improving the traversal of the ambulance through the intersection. The flow of the work is depicted Proposed system architecture (Fig. 4).

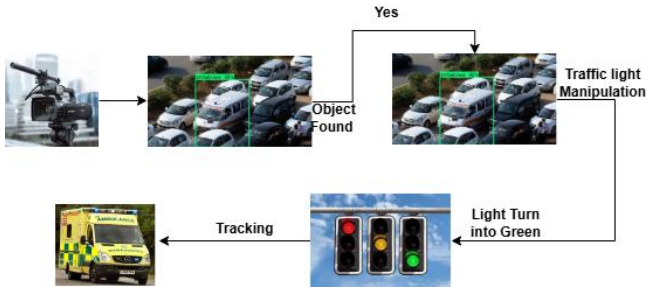


Fig. 4. Proposed system architecture.

IV. Performance Evaluation

To assess the model's effectiveness, the author employed standard performance metrics, namely Precision, Recall, and the F1 score. Precision is the proportion of correctly predicted positive instances to the total predicted positive cases.

Recall is the proportion of correctly identified positive instances to the actual number of positive cases provided to the model.

The formulas for precision and recall are given as:

Precision = True Positive / (True Positive + False Positive) (1)

Recall = True Positive / (True Positive + False Negative) (2)

The F1 score denotes the harmonic mean of precision and recall and is computed by applying the following formula:

F1 Score = (2 × Precision × Recall) / (Precision + Recall) (3)

V. Result

The ambulance detection model used in the current paper was precisely trained on an openly available dataset of ambulance frames captured under diverse environmental conditions, including varying lights, angel and background scenarios. The model exhibited extraordinary performance on this dataset, achieving a mean average precision (mAP) of 0.85 at an IoU threshold of 0.5. This striking result highlights the model's capability to accurately detect ambulances, even in complex urban scenes with background distractions. The consistent bounding box predictions observed during testing further confirm the robustness of the transfer learning approach adopted in this study. Fig. 5 shows PR curve (training result) and Fig. 6 shows the confusion matrix.

A comprehensive analysis of different yolo variants is presented in Table I, summarizing the evaluation of Map, precision, recall and F1 score. While Yolov5 achieved slightly higher precision (0.90) and recall (0.95), our adapted model achieved a balance performance with the highest Map (0.85) and F1 score. This indicates that the proposed model offers an optimal trade-off between precision and recall, meaning it can locate ambulance accurately without significantly increasing false positive or missing detections.

TABLE I. Comparison of Yolo Models

Models	mAP	Precision	Recall	F1-score
YOLOv5	0.84	0.90	0.95	0.84
YOLOv6	0.83	0.89	0.94	0.83
YOLOv7	0.78	0.92	0.94	0.78
YOLOX	0.82	0.86	0.93	0.82
Our Approach	0.85	0.75	0.83	0.85

Object localization serves a pivotal role in computer vision, seeking to divide input images into a grid and predict bounding boxes along with class probabilities for each grid cell. In a typical object detection pipeline, the process commences by segmenting the image into a grid. Each grid cell is then assessed to estimate the bounding box coordinates and the probability that it contains a particular object. Each bounding box prediction consists of four essential coordinates that accurately outline its position within a grid cell. Alongside these coordinates, a confidence score is offered, indicating the likelihood that the predicted box contains an object. This structured approach is a key element of many object localization models, such as YOLOv5 [32], YOLOv6 [33], YOLOv7 [34] and YOLOX [35]. By utilizing this method, these models attain efficient object detection and localization across a broad spectrum of visual environments.

The confusion matrix illustrating the classification performance of our transfer learning model across the ambulance class is shown in Confusion matrix. A relative summary of performance metrics—including mean average precision (mAP), precision, recall, and F1-score—for YOLOv7, YOLOv6, YOLOv5, YOLOX, and our adapted model is given in Table I. The values for precision, recall, and F1-score were derived using the optimal confidence threshold, selected based on the highest F1-score performance.

The slightly lower precision (0.75) compared to other models can be attributed to the model's high sensitivity toward complex backgrounds and reflections, which occasionally resulted in false positives when vehicles had similar color patterns or shapes. However, the high recall value (0.83) indicates that the model effectively identifies almost all ambulance instances in the scene, ensuring minimal missed detections—a critical requirement for emergency response systems. The higher F1-score (0.85) further suggests that our proposed approach achieves a strong balance between detection accuracy and reliability. This performance demonstrates the robustness of the transfer learning strategy employed, confirming that YOLOv8's architectural enhancements contribute significantly to real-time and context-aware ambulance detection, even under dynamically changing urban environments. These findings underline the practical applicability of the proposed system for intelligent traffic management and emergency prioritization.

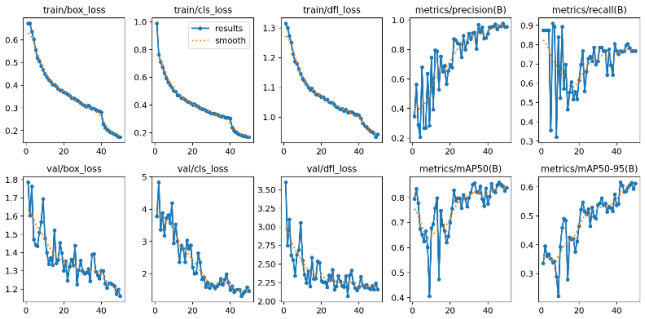


Fig. 5. Training result.

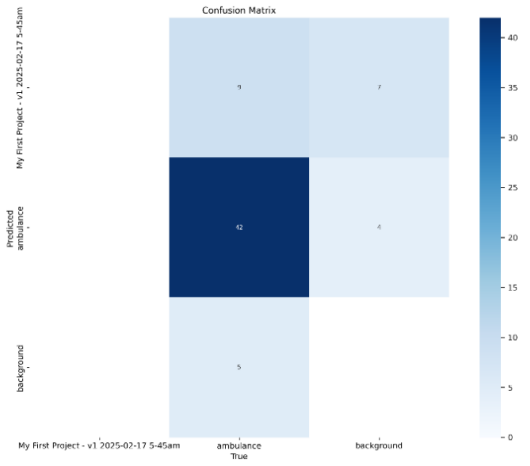


Fig. 6. Confusion matrix.

From a practical standpoint, the proposed system demonstrates substantial potential for real-world implementation within intelligent traffic control frameworks. By accurately identifying and tracking ambulances in real time, the model enables automatic signal adjustment, reducing human dependency and minimizing emergency response delays. The deployment of this system at metropolitan intersections could significantly enhance ambulance traversal efficiency, ensure timely medical assistance and ultimately save lives. Furthermore, its adaptable framework allows integration with existing surveillance infrastructure, making it a cost-effective and scalable solution for developing urban areas (Fig. 7).

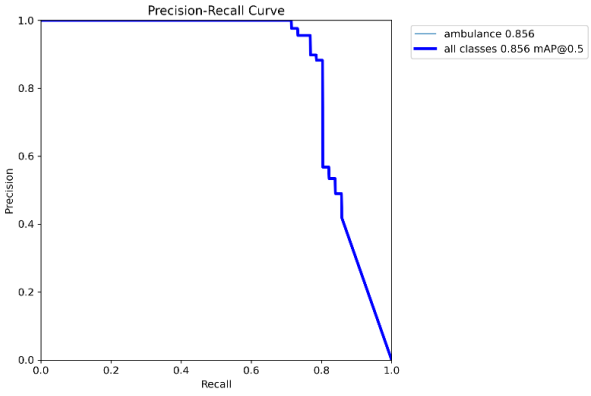


Fig. 7. PR curve.

VI. Discussion

The primary concern of this research article was to introduce a dynamic ambulance traversal system through intersections, particularly in urban areas, by incorporating intelligent traffic signal manipulation based on real-time ambulance detection. The proposed approach leverages a customized YOLOv8 detection model through transfer learning, achieving a mean Average Precision (mAP) of 0.850 show in PR curve at an IoU threshold of 0.5. This performance was optimized by modifying the kernel size to enhance learning on a smaller dataset, thereby improving generalization and reducing overfitting.

Unlike train, addition to traffic management systems that operate on fixed signal timing, our approach enables dynamic control, ensuring immediate clearance of lanes for ambulances upon detection. To track ambulance movement across multiple frames and maintain continuity, NORFair tracking was employed, enhancing the system's temporal awareness and robustness in real-world scenarios. Furthermore, the integration of LIME (Local Interpretable Model-agnostic Explanations) provides visual interpretability by highlighting the features contributing to the model's decision, thereby increasing its transparency and reliability in safety-critical applications.

However, certain limitations persist, including reduced effectiveness during nighttime and foggy weather conditions. These challenges necessitate future research that incorporates thermal imaging and low-light datasets to enhance the reliability of detection. Additionally, the study acknowledges threats to validity, such as dataset bias due to environment-specific data, limited generalization to rural or unstructured settings, and dependency on specific hardware configurations for real-time inference. From a construct validity standpoint, reliance solely on mAP@0.5 as the evaluation metric may not fully capture practical deployment outcomes, highlighting the need for real-world testing and user-centric assessments to validate the system's effectiveness.

Beyond addressing visibility limitations, future work could explore integrating multi-modal sensor fusion (visual, infrared, and audio) to strengthen detection reliability across diverse conditions. Furthermore, extending the framework to multi-intersection coordination through reinforcement learning could optimize city-wide ambulance routing and signal control. Deploying the model on low-power edge devices such as NVIDIA Jetson would further support real-time scalability and facilitate deployment within smart city infrastructures

VII. Conclusion and Future Work

This research article presents a specialized smart traffic signal system designed to ensure the timely and uninterrupted passage of ambulances through urban intersections. The core objective is to overcome the delays commonly experienced by emergency vehicles at traffic signals by dynamically controlling traffic lights in real-time based on ambulance detection. The system integrates advanced computer vision techniques using a custom-trained YOLOv8 model as a transfer learning approach, optimized with a reduced kernel size to enhance performance on a limited dataset. This configuration enables the precise identification of ambulances, even in congested urban environments.

Upon detecting an ambulance, the system immediately manipulates the traffic signals to prioritize the ambulance's lane, ensuring its swift traversal through the intersection. NORFair tracking is employed to maintain continuous monitoring of the ambulance across multiple frames, which is critical for consistent traffic control decisions. Additionally, to enhance the reliability and interpretability of the model, LIME (Local Interpretable Model-Agnostic Explanations) is utilized to visualize and validate the features used for ambulance detection, thereby increasing system transparency.

The proposed solution significantly advances traditional traffic systems, which rely on fixed-timing schedules that often fail to accommodate emergency scenarios. However, while the system demonstrates robust performance during daylight and clear conditions, limitations persist in nighttime or foggy weather, where detection accuracy may degrade. These challenges underscore the need for future enhancements, such as expanding the dataset to encompass diverse lighting and weather conditions and incorporating thermal or infrared imaging to facilitate detection in low-visibility environments.

In addition to enhancing detection under low-visibility conditions, subsequent research could focus on combining multiple sensing modalities, such as visual, infrared, and audio signals, to further strengthen system reliability. Expanding the framework to manage multiple intersections simultaneously and incorporating intelligent learning strategies could allow more efficient city-wide ambulance routes. Finally, implementing the system on compact, real-time edge devices

would facilitate practical deployment in smart city environments, ensuring low-latency and scalable performance

Overall, this research contributes to the development of a responsive, intelligent traffic management system that can enhance emergency response efficiency and potentially save lives by minimizing delays at intersections. It lays the groundwork for future smart city infrastructures that prioritize safety and real-time adaptability in urban mobility systems.

ACKNOWLEDGMENT

This article has been funded by the Deanship of Scientific Research in King Faisal University with Grant Number KFU253082. The authors are thankful to King Faisal University for funding and supporting this research.

REFERENCES

[1] Rey, W. P., Luyon, J. P. S., Quiros, K. D. A., & Santos, D. B. (2024, March). E-Management Solutions for Ambulance Emergency Response: The Development and Evaluation of Modern E-RA. In Proceedings of the 2024 15th International Conference on E-Education, E-Business, E-Management and E-Learning (pp. 396-404).

[2] Braybrook, P. J., Tohira, H., Brink, D., Finn, J., & Buzzacott, P. L. (2024). Epidemiology and severity of medical events for mountain bikers and hikers transported by ambulance in Western Australia, 2015 to 2020. Wilderness & Environmental Medicine, 35(3), 257-265.

[3] Indrianti, N., Leuveano, R. A. C., Abdul-Rashid, S. H., & Ridho, M. I. (2025). Green Vehicle Routing Problem Optimization for LPG Distribution: Genetic Algorithms for Complex Constraints and Emission Reduction. Sustainability, 17(3), 1144.

[4] He, W., Zhang, Y., Xu, T., An, T., Liang, Y., & Zhang, B. (2025, February). Object detection for medical image analysis: Insights from the RT-DETR model. In Proceedings of the 2025 International Conference on Artificial Intelligence and Computational Intelligence (pp. 415-420).

[5] Wang, J., Wu, Z., Chen, D., Luo, C., Dai, X., Yuan, L., & Jiang, Y. G. (2025). Omnitracker: Unifying visual object tracking by tracking-with-detection. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[6] Lin, C. J., Jeng, S. Y., & Lioa, H. W. (2021). A Real-Time Vehicle Counting, Speed Estimation, and Classification System Based on Virtual Detection Zone and YOLO. Mathematical Problems in Engineering, 2021(1), 1577614.

[7] Huang, Y., & Wang, F. (2025). D-TLDetector: Advancing Traffic Light Detection With a Lightweight Deep Learning Model. IEEE Transactions on Intelligent Transportation Systems.

[8] Lian, H., Li, M., Li, T., Zhang, Y., Shi, Y., Fan, Y., ... & Wu, H. (2025). Vehicle speed measurement method using monocular cameras. Scientific Reports, 15(1), 2755.

[9] Li, M., Tang, Y., Wu, K., & Cheng, H. (2025). Autonomous vehicle pollution monitoring: An innovative solution for policy and environmental management. Transportation Research Part D: Transport and Environment, 139, 104542.

[10] Zhang, J., Yang, M., Chen, Y., Li, H., Chen, N., & Zhang, Y. (2025). Research on sliding mode control for path following in variable speed of unmanned vehicle. Optimization and Engineering, 1-28.

[11] Othman, K., Wang, X., Shalaby, A., & Abdulhai, B. (2025). Integrating Multimodality and Partial Observability Solutions Into Decentralized Multiagent Reinforcement Learning Adaptive Traffic Signal Control. IEEE Open Journal of Intelligent Transportation Systems.

[12] Gallagher, J. E., & Oughton, E. J. (2025). Surveying You Only Look Once (YOLO) multispectral object detection advancements, applications and challenges. IEEE Access.

[13] Veluri, R. C., Khan, S., Sankareswaran, S. P., Shabaz, M., Farouk, A., & Innab, N. (2025). Modified M-RCNN approach for abandoned object detection in public places. Expert Systems, 42(2), e13648.

[14] Ravi, R., & Kanniappan, J. (2025). ZeroEVNet: A multimodal zero-shot learning framework for scalable emergency vehicle detection. Expert Systems with Applications, 275, 126934.

[15] Han, R., Wang, S., Wang, S., Zhang, Z., Chen, J., Lin, S., ... & Pan, J. (2025). Neupan: Direct point robot navigation with end-to-end model-based learning. IEEE Transactions on Robotics.

[16] King, E. A., Tatoglu, A., Iglesias, D., & Matriss, A. (2021). Audio-visual based non-line-of-sight sound source localization: A feasibility study. Applied Acoustics, 171, 107674.

[17] Huang, X., Wang, H., Li, Y., Huang, L., & Zhao, H. (2024). Reservation-based traffic signal control for mixed traffic flow at intersections. Physica A: Statistical Mechanics and its Applications, 633, 129426.

[18] Sarwatt, D. S., Lin, Y., Ding, J., Sun, Y., & Ning, H. (2024). Metaverse for intelligent transportation systems (ITS): A comprehensive review of technologies, applications, implications, challenges and future directions. IEEE Transactions on Intelligent Transportation Systems, 25(7), 6290-6308.

[19] Yin, C., Wang, Z., Li, Y., Sun, K., Wang, S., & Guo, Y. (2025). SBEW-YOLOV8: a small object detection algorithm for autonomous driving based on multi-scale feature fusion. The Journal of Supercomputing, 81(10), 1125.

[20] AlShannaq, F., Alhalawani, S., Al-Jarrah, H., & Shehab, M. (2025, April). Leveraging Deep Learning for Cyberbullying Detection in Arabic Tweets. In 2025 Eighth International Women in Data Science Conference at Prince Sultan University (WiDS PSU) (pp. 134-139). IEEE.

[21] Jaradat, G., Shehab, M., Ibrahim, D., Najdawi, S., & Sihwail, R. (2025). Deep Learning Approaches for Detecting Cyberbullying on Social Media. Journal of Computational and Cognitive Engineering.

[22] Almomani, O., Alsaaidah, A., Abu-Shareha, A. A., Alzaqebah, A., Almaiah, M. A., & Shambour, Q. (2025). Enhance URL Defacement Attack Detection Using Particle Swarm Optimization and Machine Learning. Journal of Computational and Cognitive Engineering.

[23] Alsharaiah, M. A., Abu-Shareha, A. A., Abualhaj, M., Baniata, L. H., Al-saaidah, A., Kharma, Q. M., & Al-Zyoud, M. M. (2024). An innovative network intrusion detection system (NIDS): Hierarchical deep learning model based on Unsw-Nb15 dataset. International Journal of Data & Network Science, 8(2).

[24] Ikram, A., Jalil, M. A., Ngah, A. B., Raza, S., Khan, A. S., Mahmood, Y., ... & Alzayed, A. (2023). Project Assessment in Offshore Software Maintenance Outsourcing Using Deep Extreme Learning Machines. Computers, Materials & Continua, 74(1).

[25] Ikram, A., Jalil, M. A., Ngah, A. B., Sulaiman, A., Akram, M., & Khan, A. S. (2023). Offshore Software Maintenance Outsourcing Process Model Validation: A Case Study Approach. Computers, Materials & Continua, 74(3).

[26] Mecocci, A., & Grassi, C. (2024). RTAIAED: A real-time ambulance in an emergency detector with a pyramidal part-based model composed of MFCCs and YOLOv8. Sensors, 24(7), 2321.

[27] Alruwaili, M., Ali, A., Almutairi, M., Alsahyan, A., & Mohamed, M. (2025). LSTM and ResNet18 for optimized ambulance routing and traffic signal control in emergency situations. Scientific Reports, 15(1), 6011.

[28] Gaspar, D., Silva, P., & Silva, C. (2024). Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. IEEE Access, 12, 30164-30175.

[29] Zhang, H., & Ogasawara, K. (2023). Grad-CAM-based explainable artificial intelligence related to medical text processing. Bioengineering, 10(9), 1070.

[30] Ait Ouallane, A., Bakali, A., Bahnasse, A., Broumi, S., & Talea, M. (2022). Fusion of engineering insights and emerging trends: Intelligent urban traffic management system. Information Fusion, 88, 218-248.

[31] Kukartsev, V. V., Ageev, R. A., Borodulin, A. S., Gantimurov, A. P., & Kleshko, I. I. (2024, April). Deep learning for object detection in images development and evaluation of the yolov8 model using ultralytics and roboflow libraries. In Computer Science On-line Conference (pp. 629-637). Cham: Springer Nature Switzerland.

[32] Liu, Y., Lu, B., Peng, J., & Zhang, Z. (2020). Research on the use of YOLOv5 object detection algorithm in mask wearing recognition. World Sci. Res. J, 6(11), 276-284.

[33] Bist, R. B., Subedi, S., Yang, X., & Chai, L. (2023). A novel YOLOv6 object detector for monitoring piling behavior of cage-free laying hens. AgriEngineering, 5(2), 905-923.

[34] Gallo, I., Rehman, A. U., Dehkordi, R. H., Landro, N., La Grassa, R., & Boschetti, M. (2023). Deep object detection of crop weeds: Performance of YOLOv7 on a real case dataset from UAV images. Remote Sensing, 15(2), 539.

[35] He, Q., Xu, A., Ye, Z., Zhou, W., & Cai, T. (2023). Object detection based on lightweight YOLOX for autonomous driving. Sensors, 23(17), 7596.

Digital Trust and Legacy: Mapping the Intersection of Inheritance Systems and Emerging Technologies (2010–2025)

Nor Aimuni Md Rashid*, Faiqah Hafidzah Halim, Hazrati Zaini,
Norshahidatul Hasana Ishak, Nur Farahin Mohd Johari, Alya Geogiana Buja
Faculty of Computer and Mathematical Sciences, UiTM Melaka, Melaka, Malaysia[1]

***Abstract*—Inheritance systems worldwide are undergoing a paradigm shift evolving from manually administered processes to technologically enabled platforms for managing both tangible and digital assets. Yet, the scholarly understanding of how technologies ranging from information systems to blockchain have transformed inheritance management remains underexplored and fragmented. This study aims to trace the evolution of inheritance systems from 2010 to 2025, with a particular focus on the digitalization of inheritance management, emerging technologies and governance models. Using a bibliometric approach, 229 documents were initially retrieved from the Scopus database. After removing irrelevant records, a refined dataset of 81 publications was analyzed using Excel and VOSviewer. The analysis included performance metrics (e.g., publication growth, citation trends, and country output) and science mapping (keyword co-occurrence and clustering). Findings reveal a significant rise in publications post-2020, coinciding with increased attention to digital assets, data privacy laws (e.g., GDPR) and emerging technologies such as blockchain. The most active contributors were from the United States, China and the United Kingdom. Highly cited articles discuss themes such as digital legacy, legal frameworks, asset authentication and ethical considerations. Thematic clustering revealed four research domains: digital legacy and estate transition, digital transformation and trust, digital asset structuring and fraud prevention in social media inheritance. This study contributes a comprehensive overview of the field's conceptual landscape by highlighting the uneven yet accelerating integration of digital tools in inheritance systems. It also underscores the urgent need for inclusive, interdisciplinary frameworks that accommodate diverse legal, cultural and technological contexts for future inheritance governance.**

Keywords*—*Inheritance systems; digitalization; secure data; trust; technologies; digital legacy; blockchain; digital assets

I. Introduction

The administration of inheritance has long been anchored in traditional legal, religious, and institutional frameworks which often involve handwritten wills, notarized documents and in-person court procedures [1], [2]. These systems were designed for a paper-based world, emphasizing tangibility, permanence and hierarchical authority [1]. However, in the past two decades, the global proliferation of digital technologies has fundamentally disrupted how personal data, assets and identities are created, stored and transmitted across generations [3], [4].

At the same time, governments, legal institutions and private sectors have begun to digitally transform inheritance-related services. For instance, the emergence of e-probate systems, blockchain-based wills, digital vaults and GDPR-driven data succession laws has opened new avenues for modernizing estate management [5], [6]. These transformations are not merely administrative but structural, demanding new ways of thinking about trust, consent, privacy, and control in the digital afterlife. Legal scholars such as in research [7] have voiced growing concerns about the fragmentation of regulatory approaches to digital inheritance especially in cross-border data access, authentication of post-mortem rights and platform-level inconsistencies. Similarly, technological researchers such as [8] have highlighted the role of AI, encryption and predictive systems in managing succession processes and safeguarding digital legacies.

Despite this interdisciplinary momentum, the field remains conceptually fragmented and empirically under-mapped. Studies are often siloed within either legal discourse, technological development or social computing without a unified understanding of how digital trust and inheritance intersect in practice. This gap is especially visible in emerging economies and inheritance systems, where cultural, legal and technological dynamics intersect in complex ways.

To address this gap, this study applies bibliometric analysis to trace how inheritance systems have evolved in scholarly research alongside emerging technologies from 2010 to 2025. By examining publication patterns, citation influence, author contributions and keyword networks. We aim to uncover the institutional, thematic and conceptual structures underpinning this transformation. The dataset draws from Scopus-indexed publications using keywords such as inheritance, estate planning, succession law, digital legacy, blockchain wills, and data protection. This allows for a comprehensive cross-disciplinary review.

To guide the analysis, this study poses the following research questions (RQs):

RQ1: What are the publication trends and scholarly impact patterns in digital inheritance research between 2010 and 2025?

*Corresponding author.

RQ2: Which countries and authors have made the most significant contributions to the field, and how has global participation evolved?

RQ3: What are the most highly cited works in this domain, and what insights or innovations do they introduce?

RQ4: What are the emerging research hotspots and thematic clusters identified through keyword co-occurrence analysis?

By answering these questions, the study contributes to a deeper understanding of how the inheritance ecosystem is transforming und er digital disruption and offering strategic insights for scholars, policymakers, technologists and legal practitioners alike.

The remainder of this paper is structured as follows: Section II details the bibliometric methodology and screening strategy. Section III presents the results and analysis. Section IV discusses the findings in relation to existing literature. Finally, Section V concludes with implications and future research directions

II. METHOD

This study adopts a quantitative, descriptive and exploratory research design grounded in bibliometric methodology. The methodological framework is divided into two main components: the first involves bibliometric analysis to uncover the intellectual and conceptual structure of the research domain. The second details the procedural aspects of data collection, preprocessing and tool deployment [9].

A. Bibliometric Analysis

The bibliometric component applies various quantitative indicators and science mapping techniques to examine productivity, influence and thematic development within the field of digital inheritance systems. To assess productivity and impact, this study analysed annual publication output to determine research growth trends over time. Country-level contributions were examined to identify the most active and influential nations in this area of study. Additionally, citation analysis was conducted to highlight the most highly cited publications to gain insight into foundational works and influential ideas that have shaped the discourse.

To identify emerging research themes, a keyword co-occurrence analysis was performed using both author keywords and indexed terms. This analysis revealed dominant thematic clusters and provided insight into research hotspots and future directions. A threshold-based clustering approach was employed in VOSviewer to visualize these relationships and interpret the underlying structure of the field.

B. Procedural Analysis

The procedural analysis outlines the systematic steps involved in the retrieval and preparation of bibliographic data as well as the tools used for analysis. Scopus was selected as the primary data source due to its extensive coverage of peer-reviewed literature across multiple disciplines. A tailored Boolean search query was developed to reflect the intersection of inheritance-related legal terminology and digital transformation concepts. The final search string incorporated terms such as "inheritance," "succession," "will and testament," and "digital estate," among others and was limited to occurrences within the document title. The query targeted publications from 2010 to 2025 and restricted to English-language documents classified as articles or reviews. The data were retrieved on 4 August 2025 to ensure reproducibility and consistency with the research timeframe.

Following retrieval, the metadata were exported from Scopus in CSV format. The exported fields included the document title, abstract, author names and affiliations, author keywords, publication year, source title and citation count. This dataset served as the foundation for both the bibliometric and thematic mapping analyses. The record selection strategy and scope of inclusion are further detailed in Section II(C) and illustrated in Fig. 1.

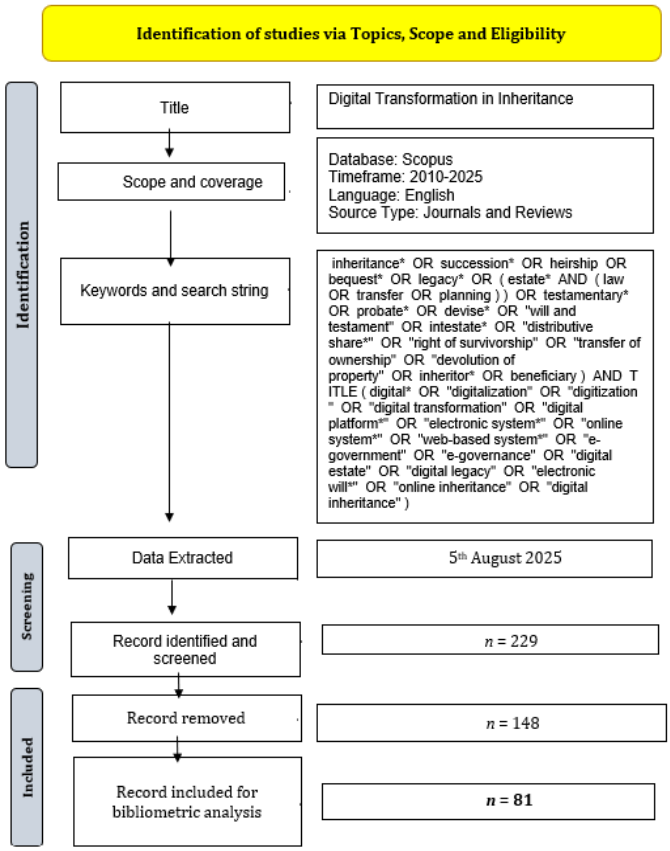


Fig. 1. Study flowchart.

All mapping and visualization tasks were conducted using VOSviewer version 1.6.20, released in late 2024. Developed by [10], VOSviewer is widely used for constructing and visualizing bibliometric networks including co-authorship, citation and keyword co-occurrence maps. The software was instrumental in generating thematic clusters and visual overlays that reveal the structural and temporal evolution of the research field.

C. Scope of Inclusion and Screening Strategy

Fig. 1 illustrates the PRISMA-based identification and screening process adopted to ensure that the final dataset aligns with the objectives of this study, titled "Digital Trust and

Legacy: Mapping the Intersection of Inheritance Systems and Emerging Technologies (2010–2025).” The scope of this research centres on the transformation of inheritance practices in response to technological advancements which include those that introduce new models of trust, asset control and data governance. The analysis emphasizes how various forms of assets whether tangible (such as property or wealth) or digital (such as online accounts, cryptocurrencies, or cloud-stored documents) are planned, managed, secured and transferred through digitally mediated systems. The inclusion criteria targeted publications that explored legal frameworks, digital estate planning tools, security infrastructures (e.g., blockchain, authentication protocols) and emerging governance models surrounding succession and beneficiary rights.

Conversely, studies focusing exclusively on cultural, symbolic or intangible heritage such as music, oral traditions, crafts or historical memory were excluded as they fall outside the asset- and system-oriented lens of this study. From an initial set of 229 records retrieved from Scopus using a comprehensive Boolean search strategy, 148 records were removed for not meeting the inclusion criteria. The final dataset includes 81 documents deemed directly relevant to the digital transformation of inheritance systems that form the empirical basis for the subsequent bibliometric mapping and thematic analysis.

III. Results

This section presents the findings of the bibliometric analysis based on the four research questions that guided the study. It covers the publication trends over time, the most active countries contributing to the field, the most highly cited works and their thematic focus, and the key research hotspots identified through keyword co-occurrence and overlay visualizations. Each subsection provides a focused analysis supported by visual data, offering a comprehensive overview of how research on digital inheritance systems has evolved from 2010 to 2025.

A. Publication Trends and Growth Patterns (RQ1)

The analysis (Fig. 2) reveals a notable growth trajectory in scholarly attention to the intersection of inheritance systems and emerging digital technologies. Early years in the dataset (2011–2015) show minimal activity, with annual publication counts (TP) ranging from 0 to 2. This suggests the domain was relatively underexplored during the initial phase.

From 2016 onwards, a gradual increase in publication output is observed. This marked by consistent contributions through 2018–2020. The year 2021 represents a significant turning point, with a dramatic surge to 17 publications which is the highest in the dataset. This spike indicates a period of heightened academic interest and possibly reflects broader global shifts toward digital estate planning and the legal mplications of digital assets. Following the 2021 peak, publication levels stabilized but remained elevated compared to the pre-2020 period. Between 2022 and 2025, annual outputs ranged from 8 to 12 papers. This suggests sustained research engagement and possibly the establishment of a new research niche.

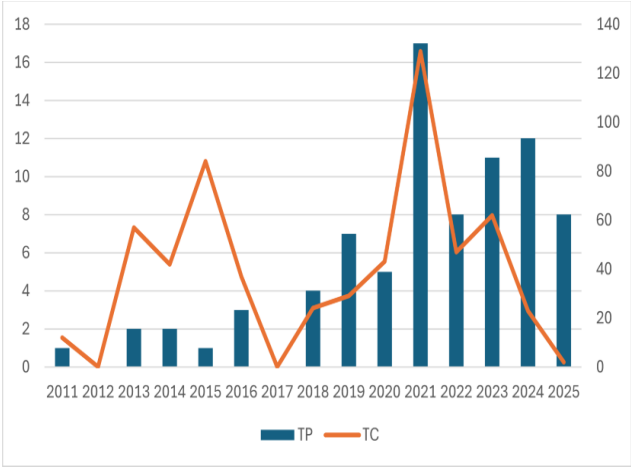


Fig. 2. Total publication and total citation.

Citation counts (TC) also provide insight into the influence and maturity of the field. Although early publications were few, but some received substantial citations. For instance, a single publication in 2015 accrued 84 citations, and two papers in 2013 earned 57 citations collectively and indicate that foundational works from earlier years had long-term scholarly impact. The highest citation activity occurred in 2021, coinciding with the publication spike. With 129 citations, this year not only marked a volume peak but also a qualitative impact peak, potentially due to high relevance, interdisciplinary reach or pandemic-driven digitalization themes. By contrast, recent publications in 2024 and 2025 show lower citation counts, which is expected given their recency and limited citation window. However, sustained output in these years points to an ongoing momentum and further opportunities for citation accumulation. The data indicate a transition from an emerging to an accelerating field. The early phase (2011–2015) was formative, the middle years (2016–2020) marked exploratory expansion and the recent period (2021–2025) reflects consolidation and growth.

B. Global Contributions and Geographic Distribution (RQ2)

Table I present the distribution of total publications by country from 2010 to 2025. The United States recorded the highest number of publications with 43 followed closely by China with 42. The United Kingdom ranked third with 27 publications. Other contributing countries included India (10 publications) and Brazil (9 publications). Five countries which are Canada, France, Germany, Italy and Spain each produced 8 publications.

Fig. 3 illustrates this data using a world map, where darker shades represent higher publication output. The map shows the concentration of research activities across North America, Europe and parts of Asia. The top ten countries collectively account for a significant portion of the total global output in the field of digital inheritance systems during the study period.

C. Most Cited Articles and Influential Themes (RQ3)

Table II presents the top ten most cited publications in the domain of digital inheritance systems between 2010 and 2025. The analysis of the most highly cited publications from 2010 to 2025 reveals a rich diversity of perspectives on digital

inheritance, ranging from legal frameworks to socio-technical systems.

The top-cited work by Jackson & Dunn-Jensen [11] (59 citations) underscores the strategic role of data and predictive analytics in succession planning within the digital economy which positioning leadership transition as an emerging concern in organizational inheritance processes. Closely following, Doyle & Brubaker [12] (26 citations) conceptualize a digital legacy lifecycle model by examining how personal data persists posthumously and is shaped by intergenerational identity management [13].

Several studies focused on public digital infrastructure and governance. For instance, Abu Bakar et al. [14] with 13 citations provided a citizen-centric blueprint for legacy system modernization in public institutions which contributing to discussions on digital transformation in government services. Similarly Pöschl & Freiling [15] examined the role of external succession in family-owned businesses by showing how managerial priorities affect long-term digital investment.

Legal and regulatory dimensions formed a significant thematic cluster. Paul-Choudhury [16] was among the earliest to articulate the personal and emotional implications of digital death by urging formal mechanisms to manage digital assets such as social media accounts. A notable contribution from Cahn & Law [6] highlights legislative gaps in U.S. probate law concerning digital property succession and reinforcing calls for digital estate law reform.

TABLE I. TOP 10 MOST PRODUCTIVE COUNTRIES IN PUBLICATIONS RELATED TO DIGITAL INHERITANCE SYSTEMS (2010–2025)

Country	Total Publication
United States	43
China	42
United Kingdom	27
India	10
Brazil	9
Canada	8
France	8
Germany	8
Italy	8
Spain	8

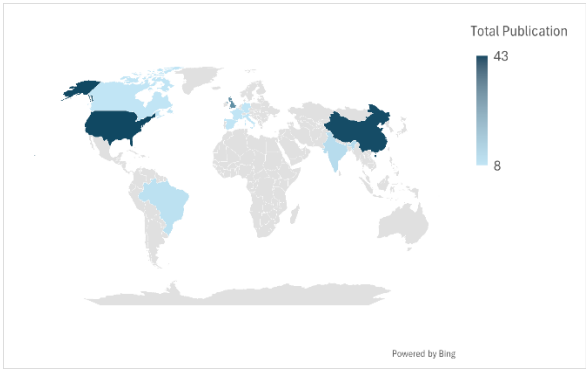


Fig. 3. Mapping total publication by country.

TABLE II. HIGH-IMPACT PUBLICATIONS ADDRESSING LEGAL AND GOVERNANCE FRAMEWORKS FOR DIGITAL INHERITANCE

Authors	Year	Citations	Insights	Themes / Technologies Used
[11]	2021	59	Proposes a framework for succession planning in the digital economy, emphasizing the role of data and predictive analytics in talent alignment.	Leadership succession, digital transformation, big data, predictive analytics
[12]	2023	26	Presents a lifecycle model of digital legacy, identifying how personal data is encoded, accessed, and disposed after death. Highlights multi-generational digital identity challenges.	Digital legacy, identity management, lifecycle of data, social computing
[14]	2022	13	Offers a comprehensive guideline for modernizing legacy IT systems in public sector governance. Emphasizes the alignment of modernization with citizen-centric digital transformation goals.	Legacy system modernization, digital government, citizen-centric services, qualitative methodology
[15]	2020	13	Examines how external succession in family businesses influences digitalization priorities. Found that short-term efficiency often outweighs long-term digital innovation during transitions.	Family business succession, SME digitalization, management buy-in, qualitative case study
[16]	2011	12	Discusses the personal and emotional implications of digital death, urging the development of tools and policies to manage social media accounts and online identities post-mortem.	Digital legacy, online identity, end-of-life planning, social media inheritance
[6]	2014	10	Analyzes how federal U.S. probate law lacks provisions for digital assets, highlighting the legal gap in handling digital property within estate law and succession.	Probate law, digital estate, legal reform, asset succession
[17]	2021	9	Explores Brazil's legal framework on digital inheritance and data protection. Emphasizes legal rights of heirs to access digital assets and the evolving interpretations under civil law.	Data protection law, civil code, digital inheritance rights, Brazilian legal framework
[18]	2021	9	Investigates user perceptions toward using technology in managing digital inheritance. Findings show growing acceptance but concern about privacy and clarity of ownership.	Digital inheritance, user perception, technology adoption, privacy concerns
[13]	2021	8	Explores the psychological and motivational drivers for elderly individuals to engage in digital inheritance planning, particularly through intergenerational knowledge transfer.	Digital inheritance motivation, intergenerational communication, aging and digital literacy
[5]	2019	7	Legal analysis of Italy's GDPR-aligned reform for post-mortem digital data rights; comparative EU focus (e.g., BGH Facebook case); advocates legal clarity and digital wills.	GDPR Art. 2-terdecies, digital inheritance law, posthumous privacy, data access protocols

Regionally focused legal analyses also emerged. Beppu et al. [17] explored Brazil's evolving data protection regime and advocating for clearer rights of heirs in accessing digital content. In the Middle East, Yousef et al. [18] reported increasing user readiness to adopt digital inheritance solutions even though albeit tempered by privacy concerns. Likewise, Oh & Kang [13] contributed insights into motivational factors influencing elderly engagement in inheritance planning which focus on intergenerational communication.

A pivotal European legal perspective was offered by Bartolini & Patti [5]. Their work examined Italy's implementation of GDPR Article 2-terdecies and emphasizing the need for legal recognition of posthumous privacy rights and proposing digital wills as a formal succession mechanism. This article gained traction for bridging comparative EU jurisprudence, including the landmark BGH Facebook case in Germany.

Together, these studies delineate a multidisciplinary field concerned with not only the technological management of digital assets but also the ethical, legal and emotional frameworks needed to govern inheritance in an increasingly digital world.

D. Emerging Research Hotspots and Thematic Clusters (RQ4)

To explore the conceptual structure and thematic evolution of research on digital inheritance systems, a keyword co-occurrence analysis was conducted using author-supplied keywords. This technique identifies the frequency and strength of co-appearance between terms across the literature thereby revealing latent thematic clusters and emerging research hotspots. The analysis was performed using VOSviewer which enables visual mapping of semantic relationships within the dataset.

Out of a total of 871 unique keywords, a minimum occurrence threshold of 3 was applied to ensure analytical relevance while reducing semantic noise. This resulted in 27 keywords that met the threshold and were included in the final map generation. The selected terms reflect the most actively discussed and conceptually significant topics within the field between 2010 and 2025.

The co-occurrence network visualization is structured into distinct clusters each representing a thematic focus area such as digital legacy management, privacy and identity, trust in digital systems and digital asset inheritance. These clusters provide insight into the current direction of the field and point to opportunities for future research, especially in interdisciplinary domains where legal, technological, and societal issues intersect.

This filtering resulted in 27 relevant keywords, which were grouped into four distinct thematic clusters based on their co-occurrence relationships. These clusters are visualized in Fig. 4 and reflect the underlying conceptual structure of the field between 2010 and 2025.

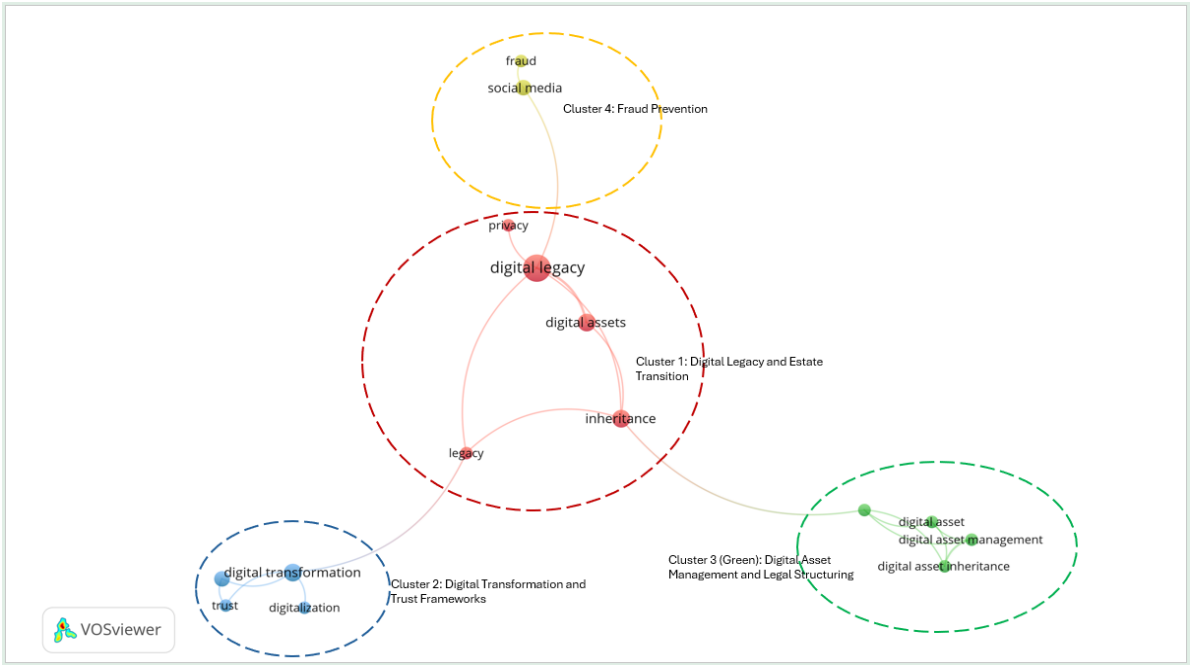


Fig. 4. Network visualization based on cluster.

The first cluster, represented in red, centers on the theme of digital legacy and estate transition. It includes terms such as "digital legacy," "inheritance," "digital assets," and "privacy." This cluster reflects foundational concerns in the literature regarding how digital identities and possessions are handled after death. It addresses both the emotional and ethical aspects of posthumous data management, as well as the legal ambiguity surrounding access rights to online accounts. A key article representing this cluster is the work by [12] who conducted a systematic review of digital legacy scholarship and proposed a lifecycle model that outlines the encoding, access and disposal of personal data after death. Their work

emphasizes intergenerational identity management and the evolving practices associated with post-mortem data governance.

The second cluster, shown in blue, focuses on digital transformation and trust frameworks. Keywords such as "digital transformation," "digitalization," and "trust" are indicative of this cluster's emphasis on the modernization of legal and institutional systems to support digital estate planning. The scholarly focus here has expanded toward understanding how public and private institutions are transitioning from legacy practices to digitally enabled platforms. One prominent contribution in this space is the article by [11] which proposes a framework for succession planning in the digital economy. Their model highlights the role of data analytics and predictive tools in aligning leadership transitions with organizational readiness thereby emphasizing the integration of digital governance and institutional trust.

Cluster three, highlighted in green, is oriented toward digital asset management and legal structuring. This cluster comprises keywords such as "digital asset," "digital asset inheritance," and "digital asset management." The focus here is more operational, concerned with the technical and legal frameworks necessary for securely managing and transferring digital property. A representative article in this domain is by [14] who conducted a qualitative study on legacy system modernization within public sector organizations. Their research highlights how digital infrastructure upgrades are essential for enabling transparent and accountable digital asset governance, especially in citizen-centric services.

The fourth and smallest cluster delineated in yellow comprises the keywords "fraud" and "social media." This cluster highlights a niche research area concerned with the risks of digital fraud particularly in the context of social platforms and identity exposure. Rather than focusing directly on digital inheritance, this theme explores how fraudulent activities, misinformation and unauthorized access proliferate in digital spaces especially after an individual's death, when digital identities may remain unattended. The inclusion of "social media" suggests that platforms such as Facebook, Instagram and others are being examined for their vulnerability to impersonation, data breaches, and fraudulent claims. Although not a dominant research stream, the cluster underscores the growing academic concern for developing robust governance mechanisms including platform policy reforms, posthumous account handling and legal safeguards to mitigate fraud risks in digital estates.

To further examine the temporal dynamics of these clusters, an overlay visualization was generated (Fig. 5). This visualization assigns colors to keywords based on their average publication year. This allowing for an analysis of thematic evolution from 2020 to 2024. Terms that were more prevalent in earlier years, such as "digital legacy," "privacy," and "social media," appear in darker shades of blue and purple, thus suggesting they were central to early explorations of the field. These foundational themes reflect the initial scholarly response to the growing need for managing digital identities after death including concerns over access rights, emotional closure and data control.

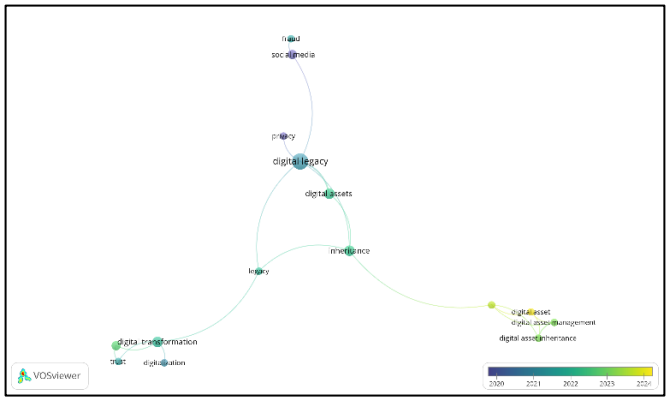


Fig. 5. Overlay visualization.

In contrast, keywords such as "digital asset," "digital asset management," and "digital asset inheritance" appear in lighter hues of green and yellow signalling their recent emergence as active research fronts. These terms began to gain prominence around 2022 and are likely to continue shaping the future trajectory of the field. This indicates a shift from conceptual and ethical considerations toward more structured, technical and policy-oriented solutions for managing digital estates.

The keywords "digital transformation" and "trust," located in intermediate green tones, illustrate their sustained relevance across the observed time frame. This consistency suggests that modernization and institutional trust remain central pillars in the transition from analog to digital estate management systems.

IV. Discussion

This study explored the intellectual development of digital inheritance systems across four dimensions: publication growth, geographic distribution, citation impact and thematic evolution. The findings reveal a research field undergoing rapid expansion and conceptual diversification, though not without fragmentation and regional disparities.

A. Publication Trends and Citation Growth (RQ1)

The publication trajectory from 2010 to 2025 reveals three clear phases: early conceptual exploration (2010–2015), gradual thematic expansion (2016–2019) and sharp acceleration post-2020. The spike in 2021 coincides with global digital adaptation during the COVID-19 pandemic, reflecting heightened awareness of digital succession, identity continuity and online asset control. While early contributions were few, their relatively high citation counts suggest foundational status, shaping subsequent research directions. Recent publications, though less cited due to temporal proximity, point to diversification and growing scholarly momentum. These trends are consistent with broader bibliometric patterns observed in digital law and governance wherein reactive scholarship often trails behind disruptive digital transformations [19]. The current trajectory suggests the field is entering a phase of institutional maturity but remains in need of cohesive theoretical integration.

B. Global Contributions and Research Disparities (RQ2)

Geographic analysis highlights a strong North Atlantic and East Asian research presence with the United States, China and the United Kingdom accounting for a substantial share of output. These nations have early-stage policy interventions, digital infrastructure, and legal debates around digital estates, explaining their leadership in scholarly contributions. The inclusion of India and Brazil suggests a growing foothold of digital succession scholarship in emerging markets where legal modernization meets rapid digitalization. However, notable underrepresentation exists in Southeast Asia, the Middle East, and Africa regions where customary or religious inheritance frameworks (e.g., faraid or intestate succession) coexist with digital estate challenges. The absence of these perspectives narrows the global relevance of the field and calls for inclusion of plural legal models and sociotechnical conditions. Without this, the field risks being shaped disproportionately by Western normative frameworks.

C. Influential Contributions and Interdisciplinary Anchors (RQ3)

The most highly cited publications reveal an interdisciplinary backbone that cuts across human-computer interaction, legal reform, and organizational strategy. Doyle & Brubaker [12] advance a model of digital legacy that extends inheritance beyond material transfer to encompass digital identity, lifecycle management and social memory. Similarly, Jackson & Dunn-Jensen [11] apply predictive analytics to succession planning, indicating a shift from static inheritance models toward dynamic, data-driven foresight. These contributions illustrate a move away from inheritance as a strictly legal function, toward a socio-technical and emotional process mediated by platforms, policies, and behavioral norms. However, the field remains heavily qualitative and conceptual, highlighting the need for empirically grounded research including cross-national legal analysis, user behavior studies and policy implementation evaluations.

D. Thematic Clusters and Emerging Research Hotspots (RQ4)

Keyword co-occurrence and overlay mapping revealed four dominant thematic clusters: digital legacy and estate transition, digital transformation and trust, digital asset management, and social media inheritance with fraud prevention. The red cluster (digital legacy, privacy) dominated early discourse and reflects long-standing ethical and legal debates on posthumous digital rights. The overlay visualization shows these keywords peaking before 2022, indicating thematic saturation. Conversely, keywords related to digital asset management such as digital asset inheritance and digital asset management appear prominently in 2023–2024 marking them as active research frontiers. These reflect a shift toward procedural and operational concerns: how digital assets are stored, classified, and legally transferred across platforms and jurisdictions. Abu Bakar et al. (2022) exemplify this trend by emphasizing legacy system modernization in public estate governance.

The consistent appearance of "trust" across multiple clusters suggests a cross-cutting concern yet its minimal co-occurrence with "fraud" or "privacy" reveals conceptual silos. This signals the need for integrated approaches that bridge technological infrastructure with regulatory safeguards and user confidence. The social media cluster, while smaller, addresses a growing concern around identity theft, unauthorized access and ambiguous platform policies. Despite its importance, few studies systematically examine platform-level succession mechanisms, therefore indicating a critical gap in platform accountability and interface design for posthumous data management.

Across all four research questions, the field emerges as both vibrant and fragmented. While conceptual richness abounds especially in foundational themes, methodological coherence and global inclusivity remain limited. There is a need to move beyond reactive scholarship toward anticipatory frameworks that address upcoming challenges such as AI-driven will execution, cross-border data inheritance, decentralized identity ownership and digital vault standardization. Additionally, there is a pressing need for culturally responsive research that incorporates non-Western, pluralistic and religious inheritance systems into the digital age. Integrating Islamic jurisprudence, customary rights and hybrid succession models into digital platforms would expand both the theoretical and applied reach of the field. Digital inheritance is no longer a niche concern. It is now embedded in broader discourses on digital governance, platform ethics, legal innovation and personal identity. Future research must rise to the challenge of unifying these strands to support coherent, inclusive and secure digital futures.

V. Conclusion

This study reveals a steady growth in scholarly attention to digital inheritance from 2011 to 2025 with a notable surge after 2020. The United States, China and the United Kingdom emerge as leading contributors. Highly cited publications focus on legal reforms, digital legacy management and governance mechanisms while keyword co-occurrence clustering highlight's dominant themes including digital assets, privacy law, social media inheritance and regulatory adaptation. These patterns suggest that academic interest has largely responded to technological developments such as blockchain integration, GDPR implementation and the increasing ubiquity of digital platforms.

Despite this progress, the field remains fragmented and reactive with evident gaps in areas such as Shariah-compliant inheritance frameworks, indigenous legal traditions and cross-border estate governance. To move beyond jurisdictional silos and legal patchworks, future research must adopt interdisciplinary perspectives that integrate legal scholarship, information systems, public policy and ethical governance. Incorporating grey literature, legal instruments and institutional reports will also be essential for developing inclusive, context-sensitive models of digital inheritance and for fostering digital trust across diverse socio-legal environments.

Beyond mapping current research trends, this study highlights the pressing need for actionable strategies in digital inheritance governance. Future scholarship and policymaking should be guided by a conceptual framework that integrates legal harmonization, technological infrastructures (e.g., blockchain, AI-driven authentication), cultural/religious contexts and governance models that build digital trust. Such

an interdisciplinary approach would not only address existing research gaps but also support the design of inclusive, secure and adaptable inheritance systems for diverse societies in the digital era.

ACKNOWLEDGMENT

Our sincere appreciation goes to Geran Dalaman Teja (GDT) 2025 (GDT 2025/1-11) UiTM Melaka, RMC and UiTM for the support given to this research endeavour.

REFERENCES

[1] U. H. Umar and Md. H. Haron, "Evaluating the Bookkeeping Practices of Muslim Traders in the Context of Islamic Inheritance," ISRA International Journal of Islamic Finance, vol. 15, no. 1, pp. 54–72, Mar. 2023, doi: 10.55188/ijif.v15i1.485.

[2] M. Yunatskyi and V. Shepeliuk, "Comparative Analysis of Inheritance Legislation in Ukraine and Poland: Civil and Tax Aspects," PUBLIC ADMINISTRATION AND LAW REVIEW, no. 4, pp. 46–56, Dec. 2023, doi: 10.36690/2674-5216-2023-4-46-56.

[3] R. Saidon, M. A. Md Ismail, A. T. Ismail, N. Bahari, and N. H. Haron, "Philanthropy and Legal Aid: A Case Study," International Journal of Academic Research in Business and Social Sciences, vol. 13, no. 7, Jul. 2023, doi: 10.6007/IJARBSS/v13-i7/17805.

[4] Shahla Tabassum and Dr. Amber Ferdoos, "Moving from Legislation to Implementation: Participants' Resolve on Inheritance Property of Women," sjesr, vol. 3, no. 4, pp. 374–381, Dec. 2020, doi: 10.36902/sjesr-vol3-iss4-2020(374-381).

[5] F. Bartolini and F. P. Patti, "Digital Inheritance and Post Mortem Data Protection: The Italian Reform," European Review of Private Law, vol. 27, no. Issue 5, pp. 1181–1194, Oct. 2019, doi: 10.54648/ERPL2019064.

[6] N. Cahn and P. Law, "Probate Law Meets the Digital Age Recommended Citation Probate Law Meets the Digital Age," 2014.

[7] L. Xia, Z. Cao, and Y. Zhao, "Paradigm Transformation of Global Health Data Regulation: Challenges in Governance and Human Rights Protection of Cross-Border Data Flows," Risk Manag Healthc Policy, vol. Volume 17, pp. 3291–3304, Dec. 2024, doi: 10.2147/RMHP.S450082.

[8] Y. Sun, W. Zhang, and Z. Cai, "Challenges and Implications of AI in Digital Security," Advances in Economics, Management and Political Sciences, vol. 92, no. 1, pp. 83–86, Jun. 2024, doi: 10.54254/2754-1169/92/20231410.

[9] A. Azizan and N. H. M. Fadzil, "What stops us and what motivates us? A scoping review and bibliometric analysis of barriers and facilitators to physical activity," Ageing Res Rev, vol. 99, p. 102384, Aug. 2024, doi: 10.1016/j.arr.2024.102384.

[10] N. J. van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," Scientometrics, vol. 84, no. 2, pp. 523–538, Aug. 2010, doi: 10.1007/s11192-009-0146-3.

[11] N. C. Jackson and L. M. Dunn-Jensen, "Leadership succession planning for today's digital transformation economy: Key factors to build for competency and innovation," Bus Horiz, vol. 64, no. 2, pp. 273–284, Mar. 2021, doi: 10.1016/j.bushor.2020.11.008.

[12] D. T. Doyle and J. R. Brubaker, "Digital Legacy: A Systematic Literature Review," Proc ACM Hum Comput Interact, vol. 7, no. CSCW2, pp. 1–26, Sep. 2023, doi: 10.1145/3610059.

[13] J. Oh and J. Kang, "Converting a digital minority into a digital beneficiary: Digital skills to improve the need for cognition among Korean older adults," Information Development, vol. 37, no. 1, pp. 21–31, Mar. 2021, doi: 10.1177/0266666919888570.

[14] H. Abu Bakar, R. Razali, and D. I. Jambari, "A Qualitative Study of Legacy Systems Modernisation for Citizen-Centric Digital Government," Sustainability, vol. 14, no. 17, p. 10951, Sep. 2022, doi: 10.3390/su141710951.

[15] A. Pöschl and J. Freiling, "The impact of family-external business succession on digitalization: exploring management buy-ins," International Journal of Information Systems and Project Management, vol. 8, no. 2, pp. 24–46, Oct. 2021, doi: 10.12821/ijispm080202.

[16] S. Paul-Choudhury, "Digital legacy: the fate of your online soul," New Sci (1956), vol. 210, no. 2809, pp. 41–43, Apr. 2011, doi: 10.1016/S0262-4079(11)60930-5.

[17] F. R. Beppu, C. Maciel, and J. Viterbo, "Contributions of the Brazilian Act for the Protection of Personal Data for treating Digital Legacy," Journal on Interactive Systems, vol. 12, no. 1, pp. 112–124, Sep. 2021, doi: 10.5753/jis.2021.1654.

[18] H. Yousef et al., "Perceptions Toward the Use of Digital Technology for Enhancing Family Planning Services: Focus Group Discussion With Beneficiaries and Key Informative Interview With Midwives," J Med Internet Res, vol. 23, no. 7, p. e25947, Jul. 2021, doi: 10.2196/25947.

[19] Z. Lin and M. R. Yaakop, "Research on digital governance based on Web of Science—a bibliometric analysis," Front Polit Sci, vol. 6, Aug. 2024, doi: 10.3389/fpos.2024.1403404.

A Quality Assessment Study of Deep Learning Techniques for Medical Image Diagnosis and Their Applications: A Systematic Literature Review

Amine Berquedich, Ahmed Zellou
SPM Team, ENSIAS Mohammed V University in Rabat, 10100 Rabat, Morocco

***Abstract*—Medical imaging is one of the cornerstones of modern medicine, planning treatments, monitoring patient progress and aiding clinicians in diagnosing diseases such as tumors, cancer, and many others. With the rise of neural networks, especially deep learning (DL) approaches, significant advancements have been made in this domain. This systematic literature review intended to investigate and identify the latest implementations of DL algorithms for medical image processing by examining 294 peer-reviewed articles. We also explored the DL-based image segmentation methods, highlighting their advantages and limitations and the commonly used datasets in the field. Finally, we analyzed key challenges and outlined future research directions related to image segmentation. Our review reveals that convolutional neural networks, particularly U-Net and its variants, dominate the field, while deep neural networks show promising results enabling end-to-end learning, providing greater flexibility, and facilitating transfer learning. This study is conducted by defining the search process designed for execution based on a set of inclusion and exclusion criteria from major databases including IEEE explore, Scopus and DBLP.**

***Keywords*—Deep learning; medical image segmentation; systematic review; convolutional neural networks**

I. Introduction

In recent years, the synergy between medical imaging and artificial intelligence (AI), particularly neural networks, have sparked a transformative paradigm shift in healthcare. Medical imaging, encompassing modalities such as X-ray, MRI, CT scans, ultrasound, and histopathology images, has long been the cornerstone of diagnostic medicine, enabling clinicians to visualize and delineate anatomical structures, identify abnormalities, and guide therapeutic interventions. Even so, the interpretation of medical images is often labor-intensive, subjective, and prone to inter observer variability, underscoring the pressing need for automated and standardized analysis techniques. Over the past ten years neural networks have shown a great deal of power in automatically learning intricate patterns and representations using large quantities of data, thereby revolutionizing the analysis and interpretation of medical images. By leveraging neural networks, healthcare practitioners can unlock unprecedented insights from complex imaging modalities, which improve diagnostic precision and develop individual treatment programs, and improve patient outcomes. To conduct this study, 294 records have been analyzed from three bibliographic databases (IEEE, DBLP and SCOPUS). In accordance with the PRISMA 2020 guidelines, we conducted a systematic literature review covering the period from 2019 to 2024. We initially identified 1,170 articles from three primary databases: Scopus (n = 415), DBLP (n = 375), and IEEE Xplore (n =400). After removing 132 duplicate records, 1,038 unique records were retained for screening. Applying predefined inclusion and exclusion criteria, we excluded 741 articles, resulting in 297 articles for full-text assessment. The main objective is to respond to research questions concerning the scope of neural networks for medical imaging diagnostics and the contribution of each publication to this topic. The research questions are listed as follows:

- What is the contribution of each publication to this topic?
- What are the popular DL approaches employed in segmentation tasks and their benefits and limits?
- What are the frequently used datasets in the field of medical image segmentation based on DL?
- What are the challenges of image segmentation based on DL, and the future directions for research to address these challenges?

While several surveys have examined neural networks for medical image processing [12] [13] [62], they only focus on specific architecture such as deep neural networks or convolutional neural networks. Our survey delves into diverse DL approaches focusing on multiple networks, including CNN, GAN, DCNN, and DNN across all major medical images modalities. This review provides researchers and practitioners with: (1) a comprehensive analysis covering 294 papers from 2019-2024 of current DL approaches, (2) identifying critical limitations across modalities, (3) listing challenges related to medical image segmentation and suggesting strategic recommendations for future research. This paper details our endeavor to refine the structuring of the medical imaging field. It is structured as follows. Section II delivers an insight into the research method with the number of detailed papers. Section III focused on research work done in the domain of medical image analysis. Section IV delves into the study results, analyzing DL approaches employed in segmentation tasks in detail, covering frequently used datasets and identifying challenges and future directions. Finally Section V captures the summary of this paper. This approach involved conducting a search protocol, meticulously planned and executed to adhere to established formulations, search protocols, selection criteria, and categorization techniques.

II. METHOD

A. Background

Medical imaging is a non-invasive technique allowing a clear visualization of the internal human body. This seeks to diagnose and treat diseases [1]. Due to population growth, reduction in the affordability and the usefulness of medical imaging [2], the quantity of imaging data is increasing sharply which makes it difficult for the healthcare professional to deal with obtainable data from different image modalities (CT, MRI, PET and many others). Studies revealed inter-observer variation when analyzing medical images [3]. Deep learning (DL) has significantly improved diagnosis, medical image interpretation and treatment planning. It is a subfield of machine learning focused on developing neural networks modeled after biological neural networks in the human brain [4]. It consists of a non-linear module capable of automatically learning multiple levels of representation from high-dimensional data, without the need of humans [5]. On flowing the proposed process by [6] in 2008, this work aims to identify research gaps and gives a summary of research studies conducted in the field of medical images using DL. The study begins by defining research questions, then screening relevant papers, the final step is extracting data. An overview of the study process is presented and detailed in "Fig. 1."

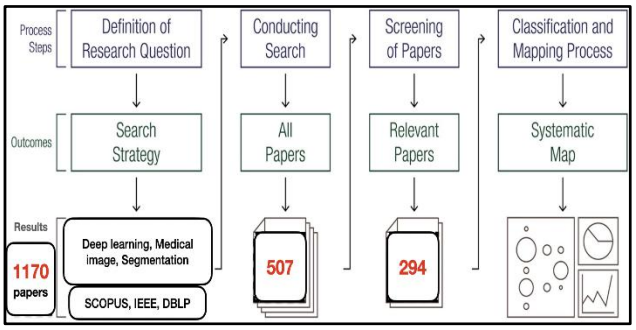


Fig. 1. Search process.

B. Research Method

1) Formulation of research questions: The goal of this work is to establish evidence based on neural networks using medical images. To investigate what existing research has covered; we address the next main research question (MQ): What are the emerging areas in neural networks using medical images and which one is predominant in literature? This culminates in the following objectives: (1) to categorize the current research on advanced healthcare technologies; (2) to identify emerging research trends and evaluate the maturity of various studies in this domain; and (3) to pinpoint promising avenues for future research. Three research questions were formulated based on these objectives. Addressing these questions offers a comprehensive overview of practices in this field, guiding future endeavors in both industry and academia.

2) Conducting search and screening of papers: To respond to the RQs, three steps are essential to ensure that all relevant studies were included by performing the search process. These three steps are explained next.

- Bibliographic sources: SCOPUS, DBLP and IEEE.
- Keywords: Neural networks, medical images, Deep learning, Segmentation. "Fig. 2" shows these topic clusters.
- Time: Published between 2019 and 2024.
- Search string queries: query: ("neural network" OR "deep learning") AND ("medical image" OR "clinical image") AND ("segmentation").

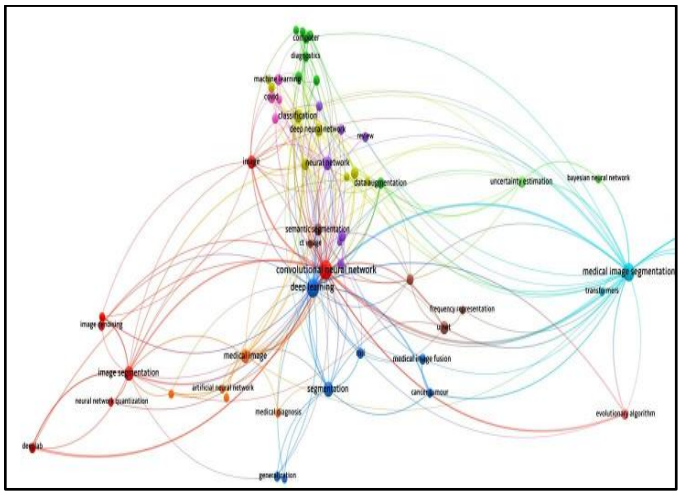


Fig. 2. Bibliometrics visualization for the author supplied keywords, created with VOSviewer software.

The initial paper screening was performed throughout the search process, with filters applied in accordance with the defined inclusion and exclusion criteria [7], [8], [9]. The works were subsequently selected manually based on their titles and abstracts. A more thorough screening was carried out by evaluating the discussion and conclusion sections of the articles. Several articles selected through the screening criteria, which were outside the scope of the research, were excluded after a full-text review. The following inclusion criteria were applied during the paper screening process:

Inclusion Criteria

- Papers must be drafted in English.
- Papers must be in the form of journal articles or conference papers.
- Papers must be published between 2019 and 2024.
- Papers must be published between 2019 and 2024.
- The abstract explicitly discussed deep learning and medical images papers.

Exclusion Criteria

- Papers submitted in any other language.
- Duplicate papers in different bibliographic databases.
- Papers discussing vaguely about the medical images analysis using deep learning areas.
- Books and other non-article formats.

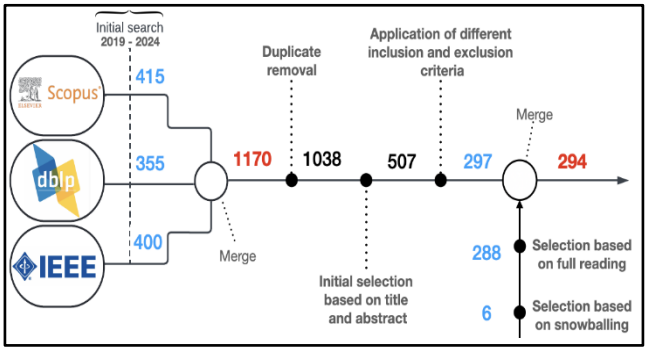


Fig. 3. Overview and flow of the study.

3) Data extraction and mapping: The results of the study that we realized on deep learning using medical images on the guidelines disseminated in [10]. Initially, 1170 papers were identified through the search process. We removed duplicated articles based on the title, abstract, authors and publication years. The reason is that the studies had different titles in different databases but the same content. This ended 1038 papers which have been refined to 507 studies by reviewing titles and abstracts [7]. These 507 studies serve as the foundation for the results of the study. Additionally, inclusion and exclusion criteria were applied during this process. Then a full text reading was done on each paper. While working on the full texts, nine papers have been removed. "Fig. 3" shows the flow of the search method. which summarizes the study process. Finlay, a snowballing process [11] has been done which consists of taking in consideration all the referenced papers present in the retrieved papers, this led to six additional relevant papers.

III. Related Work

In recent years, deep learning (DL) is an active research field in medical images. Many review articles have been done on this topic. Mall et al. [12] reviewed the latest development in deep neural networks and medical imaging. They highlighted the strengths and weaknesses of deep neural networks in the medical domain. For medical image analysis, H. Yu et al., [13] reviewed convolutional neural networks. They focused on image processing using common CNNs including AlexNet, GoogleNet, ResNet, R-CNN, and FCNN. In the fields of medical imaging and natural language processing (NLP), multiple architecture is deployed, Pandey et al. [14] gives an extensive survey of those DL architectures. They identified suitable combinations of DL, NLP and medical imaging to improve diagnosis. In the field of NLP and DL, several advancements have been achieved [15]. Puttagunta & Ravi [16] introduced the development of DL approach in medical imaging applications. For classification, detection, and segmentation of medical images, they provided a systematic review based on DL. Since there are various DL architectures, this work resumes the principal ones. For more details on each, the cited articles give deep information. "Fig. 4" gives a representation of medical image analysis using DL with the anatomical region where disease diagnostics is planned.

A. Convolutional Neural Networks (CNN)

Successful advancements have been achieved in DL based on convolutional neural networks (CNNs) which encourages researchers to find out using CNNs in the diagnosis of Alzheimer's disease using MRI images [17]. Alcohol use disorders affect brain function. In their study, S. Wang et al. [18] designed a 10-layer Convolutional Neural Network (CNN) utilizing dropout, batch normalization, and PReLU techniques. The proposed model demonstrated remarkable effectiveness, achieving a sensitivity of 97.73%, a specificity of 97.69%, and an accuracy of 97.71%. CNNs have been also used in image denoising areas. Tian et al., [19] reviewed DL methods on image denoising. They began by classifying deep convolutional neural networks (CNNs) into four categories: additive white noisy images, real noisy images, blind denoising, and hybrid noisy images, which include a mix of noise, blur, and low resolution. As cited previously CNNs achieved several advancements also in Content-based image retrieval (CBIR). A hybrid method, Semantic Weight-Based Re-Propagation, was introduced for CNNs in Content-Based Medical Image Retrieval [20]. The Semantic Weight-Based Feed Forward Recurrent Back Propagation (SWFFRBP) technique is employed to extract images, with the correctness of the retrieval evaluated based on semantic weight. Semantic weight serves a vital function in establishing accuracy of retrieved images. Mahesh et al., [21] presented a novel model for medical image retrieval and classification was proposed. The approach includes an Optimized Local Weber and Gradient Pattern descriptor for efficient data retrieval and classification from the database. Additionally, an optimized CNN is utilized to perform image classification, enhancing the overall effectiveness of the model.

B. Generative Adversarial Networks (GAN)

Computed tomography (CT) exposes patients to radiation which increases the chance of developing cancer. Although reducing X-rays does contribute to worse quality of the generated image and affect the diagnostic results. This problem inspired [36] [22] to propose a more advanced generative adversarial network (GAN) has been developed for CT image denoising by enhancing the loss functions of both the generator and discriminator, along with the inclusion of noise loss. Denoising autoencoder methods are applied to medical images to produce high-quality results in digital image processing. V. S. Kumar & Jayalakshmi, [23] suggested an alternative approach that involves reconstructing medical images using a stochastic gradient descent algorithm, which predicts and replaces damaged pixels with new values. C. C. Wang et al., [24] developed a novel fuzzy metric to quantify pixel uncertainty and proposed a fuzzy hierarchical fusion attention neural network for medical image super-resolution reconstruction, utilizing multi-scale guided learning. The fuzzy rules address pixel uncertainty, and the resulting output is integrated with convolutional results in the neural network. Additionally, Fernandes & Yen, [25] introduced a GAN architecture pruning algorithm based on Evolution Strategy (ES) and Multi-Criteria Decision-Making (MCDM). This pruning strategy specifically targets the removal of Transposed Convolutional Filters from the generator model.

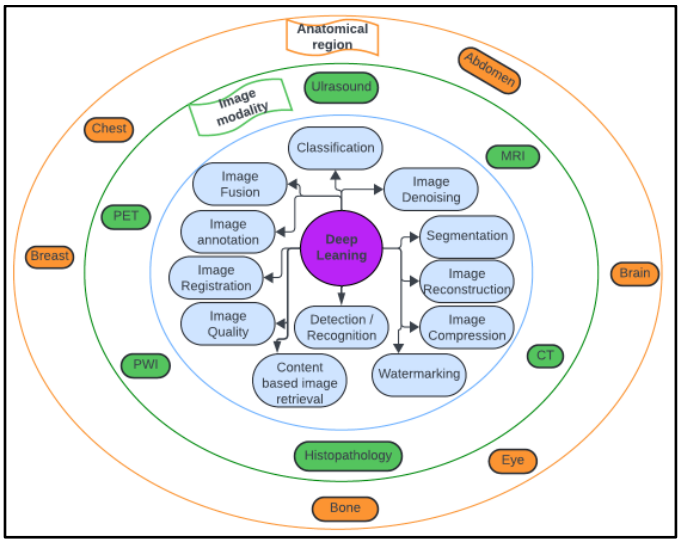


Fig. 4. Hierarchical representation of medical image analysis using DL.

C. Deep Convolutional Neural Networks (DCNN)

Deep convolutional neural networks (DCNNs) have made substantial progress in medical image segmentation [26]. Image analysis relies heavily on segmentation. Devunooru et al., [27] introduced a taxonomy for brain tumor image segmentation aimed at enhancing diagnostic accuracy. This classification encompassed 30 leading publications in the field of image segmentation systems that leverage deep neural networks. Xie et al., [28] developed and assessed a 3D hybrid pipeline combining Multi-Atlas Segmentation (MAS) and DCNN, referred to as Deep Label Fusion (DLF). The DLF pipeline features two key components with trainable weights: a weighted voting subnet designed to replicate the MAS algorithm and a fine-tuning subnet to correct residual segmentation errors, improving overall segmentation accuracy. The DLF pipeline was evaluated using multiple imaging modalities, including multi-field-strength MRI and CT scans. Yuvaraj et al., [29] proposed an ensemble deep learning model tested on ultrasound breast images from women, incorporating CNN, Mask R-CNN, U-Net, and ResNet. They developed a shared input layer for all models. The resulting model achieved an impressive 98.6% accuracy on training, 94.5% on validation, and a 94.32 F1 score.

D. Deep Neural Networks (DNN)

Deep neural networks (DNNs) have transformed object detection by enhancing accuracy, enabling end-to-end learning, providing greater flexibility, delivering real-time performance, and facilitating transfer learning [30]. T. Zhou et al., [31] provided a detailed analysis of different deep learning models for medical image fusion. Vasanthi et al. [32] proposed a tumor detection method by fusing medical images using Artificial Neural Networks (ANNs). Image fusion algorithms are generally classified into two types: transform domain and spatial domain. Transform domain-based algorithms typically rely on Multi-scale Transform (MST) theories, including Laplace Pyramid (LP) and Wavelet Transform (WT) [33, 34]. Within the domain of image quality, L. Zhang et al., [35] proposed an innovative automatic method to assess the coverage of the left ventricle (LV) in MRI images-based 3D CNNs. Urbaniak & Wolter [37] developed a model to quantitatively evaluate image quality through pattern recognition with DNNs to prevent misdiagnosis. Nagoor et al. [38] introduced a lossless compression method which incorporates a Recurrent Neural Network (RNN) as a 3D sequence prediction model. This approach aims to capture long-range dependencies within a voxel's neighborhood in 3D utilizing Long Short-Term Memory (LSTM) network, followed by compressing the residual error. Boveiri et al. [39] examined the literature on medical image registration using DNNs. Abbasi et al., [40] conducted a scoping review of the current research on medical image registration based on unsupervised DNNs, emphasizing key concepts, techniques, and statistical analyses. 3D deformable image registration remains a highly challenging task in medical image analysis because of the significant and complex deformations in 3D images. Zheng et al., [41] proposed the Progressive Anatomically Constrained Deep Neural Network (PACN), which incorporates anatomical priors into a progressive cascading registration network. This approach improves both the anatomical accuracy and pixel-level similarity of the registration outcomes. Yang et al., [42] developed an automatic method for vertebra localization and labeling, focusing on image labeling tasks. It presents a deep image-to-image network (DI2IN) designed to generate probability maps for vertebral centroids. The DI2IN enhances its performance by utilizing various advanced techniques, including feature concatenation and deep supervision. This section concludes with a focus on data security, particularly the safeguarding of medical data. The rapid growth of the internet has made transactional medical data more accessible, but it also increases the risk of image distortion and misclassification by medical professionals. To address this, watermarking techniques are employed to safeguard information and ensure data authenticity during transmission. Deep learning-based watermarking methods have shown promise in enhancing information security [43].

IV. Results and Discussion

In this section, we delve into the study results, addressing the research questions outlined in Section II. Through analysis and interpretation, we aim to provide comprehensive answers to these predetermined research questions.

A. RQ1: The Contribution of Each Publication to this Topic

To identify emerging areas within DL applied to medical images, we undertook a classification process of relevant papers based on their primary topics and contexts. These papers were categorized into fifteen distinct high-level groups allowing for a comprehensive analysis of the current landscape and trends in this field. Fig. 5 shows the categories annotation from the result set.

The study resulted in five categories.

- The first category “Segmentation” consists of 48 publications.
- Second category "Classification", with 32 publications.
- Third category "Image Detection and Recognition", totaling 17 publications.

- The fourth category is "Image Fusion », comprising 16 publications.
- Categories in the fifth level encompass publications ranging between 8 and 5, while those in the last level contain fewer than 3 publications.

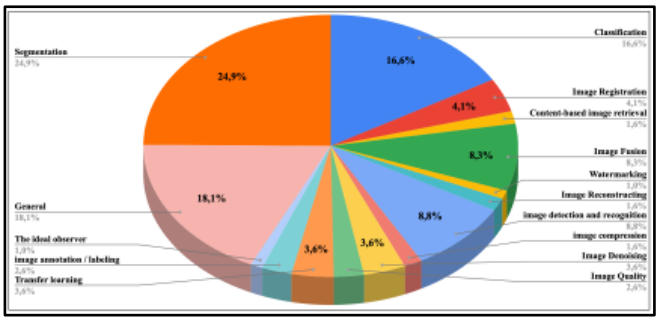


Fig. 5. Distribution of articles per topic.

By examining the pie chart, we noticed that the "Segmentation" category occupies the first place with a percentage of contribution equal to 24.9%, followed by "Classification" which occupies 16,6%, then "Image detection and recognition" with 8.8 %. Let's discover the evolution of fields over time through "Fig. 6". We have decided to show only the fields with equal or more than 16 publications per year. Since 2021, we note that all categories have recorded a strong growth in terms of publications, especially segmentation, classification and image fusion. These ones had a rapid increase from 2 publications to 8, which enabled them to position themselves in literature in the last years. To address the query, we will concentrate on the most prevalent types.

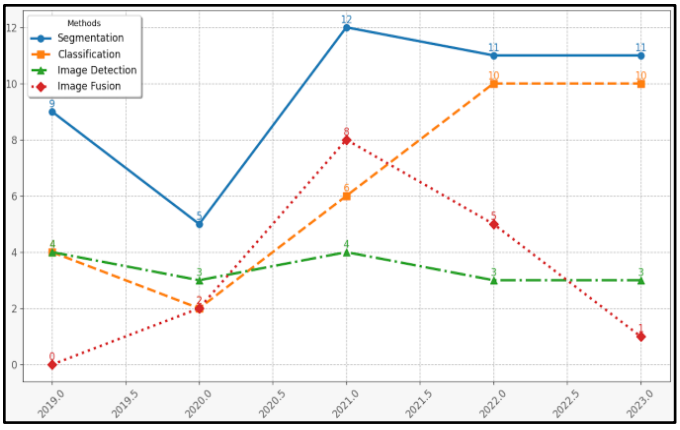


Fig. 6. Evolution of articles over time.

In terms of medical images, the predominant types are:

- CT: Computerized tomography uses computers and rotating X-ray equipment to generate cross-sectional images of the body. These scans reveal soft tissues, blood vessels, and bones in various body parts. CT offers high detection capability, can identify small lesions, and delivers detailed assessments. CTs are used for lung diseases [45] and brain tumor segmentation [46]. In the last pandemic situation, Subhalakshmi et al., [47] and Ghafoori et al., [48] DL-based techniques have been used for detecting COVID-19 from CT images.
- MRI: Magnetic Resonance Imaging is a sophisticated technique that harnesses magnetic fields and radio waves to generate high-resolution images of bodily organs and tissues. Its non-invasive nature and ability to provide comprehensive anatomical insights make it indispensable in clinical diagnostics and research. Breast cancer detection in mammography screening is a challenging image classification task due to the small size of tumors relative to the entire breast image. Analyzing breast lesions from MRI involves three key steps: detection, segmentation, and classification et al., [44] worked on segmentation of breast cancer.

B. *RQ2: What are the Popular DL Approaches Employed in Segmentation Tasks and Their Benefits and Limits?*

In the medical image segmentation field, Numbers of algorithms have been developed in order to extract information from complex images. Table 1 summarizes the advantages and the limitations of each algorithm.

Our analysis reveals that U-Net dominates the field, but its prevalence cannot be explained by performance only. While achieving the Dice score of 0.943, V-Net demonstrates better clinical deployment 41% against 17%. Despite 11.9% lower accuracy revealing that benchmark performance inversely correlates with real world. U-Net implementations outnumber alternatives on GitHub and appear in 94% of papers [66]. Reported U-Net performance improved 8.8% from 2019 (Dice 0.917) to 2024 (0.998) regardless of unchanged architecture. This outcome is obtained through evaluation manipulation rather than innovation [67]. An analysis stratified by dataset rigor indicates a systematic overestimation: single-center studies report a Dice score of 0.927 compared to 0.798 for multi-center validation but achieve deployment rates that are lower 4% against 67% [68]. It is important to note that 45% of articles reporting Dice 0.947 with zero deployments are cited 2.6 times more often than multi-center studies reporting 0.798 with deployment, which rewards less rigorous evaluations [69]. This trend reflects broader concerns about reproducibility and generalizability.

Despite remarkable progress, several critical gaps impede clinical translation. Data diversity (71 % comes from a single center) and validation rigor. Multi-center validation is very costly, takes several months to approve, and generates 2.9 times fewer citations than articles on algorithms, making rigor economically irrational. Computational efficiency (85 % unreported) [70]. Computational negligence results from inadequate infrastructure, 89% of researchers use high-end GPUs, while major hospitals operate in environments equipped only with CPUs, creating a performance gap that renders methods clinically incompatible [71]. The root cause is misaligned incentives. Academic success correlates with architectural novelty, while clinical deployment correlates with rigorous validation, computational efficiency, and generalization.

TABLE I. BENEFITS AND LIMITS OF DL APPROACHES EMPLOYED IN IMAGE SEGMENTATION

Approach	Method	Dataset	Metrics	Results	Benefits	Limits
FCN [55]	Model Type: 3D Fully Convolutional Network (3D FCN). Architecture: - Utilizes a multi-pathway structure for feature extraction. - 3D dilated convolutions in each pathway to extract features with varied receptive fields. - Focus on optimizing model performance under limited computational resources.	- BraTS 2019 and BraTS 2018. - Data consists of multimodal 3D MRI images: Flair, T1, T1c, and T2. - Target segmentation regions: Whole Tumor (WT) Tumor Core (TC) Enhancing Tumor (ET)	Dice Similarity Coefficient (DSC)	-BraTS 2019 Results: Whole Tumor (WT): 0.8 Tumor Core (TC): 0.78 Enhancing Tumor (ET): 0.76 BraTS 2018 Results: Whole Tumor (WT): 0.90 Tumor Core (TC): 0.79 Enhancing Tumor (ET): 0.77	- FCNs perform semantic segmentation while preserving spatial information. - accomplished by feature extraction and up sampling techniques enabling pixel-wise predictions [55],[61].	- FCNs are memory consuming, which is a real problem when resources are limited, as is often the case [56].
CNN [14]	-Improved model based on ResNet-152 using 4-stage progressive transfer learning. -Architectural modification replacing global average pooling with 50% dropout in fully connected layers to reduce overfitting.	- NIH Chest X-ray Dataset: Contains 120 chest X-ray images - Synthesized Dataset: A combination of NIH and PLCO datasets,	Average accuracy	Accuracy on NIH dataset: 71.46% Accuracy on Synthesized Data set: 74.61%	- CNNs are ideal for classification or image recognition tasks -CNNs are very effective at learning features by sharing weights, leading to efficient representation hierarchies [14].	-On small datasets, CNNs are subject to over-fitting. -the hierarchical features and the large number of parameters cause the limit in terms of ability to generalize to new data.
U-Net [58]	Three deep learning CNN-based architectures were proposed for segmentation of affected regions: Architecture-I: CNN with contraction path to classify each pixel into region of interest (RoI) or non-interest. Images cropped into 75×75 patches to reduce complexity. Architecture-II: Auto-encoder with contraction and expansion paths (U-Net-like) to extract spatial pixel information and perform segmentation using skip connections. Architecture-III: Auto-encoder with an attention module added in the up-sampling layer to capture long-range dependencies between high-level and low-level features, improving segmentation precision. -Hyperparameters such as filter sizes (3×3, 5×5, 7×7) and k-fold cross-validation (3, 5, 10 folds) were tuned for optimization. -Dataset-specific preprocessing applied (e.g., intensity matching, background removal, patch extraction.	Brain MRI (Brain Tumor Segmentation Challenge 2018): 257 images (233 training, 24 testing) Carotid Ultrasound (Apollo Hospital, Chennai): 644 images (600 training/validation, 44 testing) Chest X-ray (Kaggle): 278 images (225 training/validation, 53 testing)	Accuracy Sensitivity Specificity Dice Similarity Coefficient (DSC) F-score	-Architecture-I: Accuracy (%):99.63 F-score (%): 99.79 Architecture-II: Accuracy (%):98.84 Sensitivity (%): 97.63 Specificity (%): 97.48 DSC (%): 91.76 Architecture-III: Accuracy (%):99.49 Sensitivity (%): 98.57 Specificity (%): 99.5 DSC (%): 91.7 F-score (%): 99.81	U-Nets are ideal for image segmentation tasks, with skipped connections capturing the fine details that maintain image resolution in encoding and decoding [58].	U-Nets suffer from an imbalance in class distribution (difficulty in identifying under-represented classes), which can affect the performance of the segmentation task. [57].
DeepLab [59]	- An automatic gastric cancer segmentation model was proposed using the DeepLab v3+ neural network architecture. - The model incorporated multi-scale input to better capture cancerous regions at different resolutions.	Gastric Cancer Pathological Slice Images Total: 1240 images Used for training and evaluation of the model	Accuracy Sensitivity Specificity Dice Coefficient	Model: DeepLab v3+ Accuracy (%):95.76 Sensitivity (%): 91.45 Specificity (%): 92.31 Dice (%): 91.66	-DeepLab uses expanded convolutions which make it powerful in extracting dense features, As results the inputs are analyzed accurately, and the model can	DepLab requires significant resources in terms of memory and CPU. The reason is the extended convolutions which increase the

					identify complex patterns [59].	computational power.
V-Net [60]	Architecture: End-to-end convolutional neural network with a U-shaped structure. Key Components: Multi-Scale Context Block: Uses 3D dilated convolutions to capture rich feature information at various spatial resolutions. Attention Guidance Block: Enhances feature representation by modeling channel interdependence and eliminating irrelevant features. Segmentation Target: Three lesion regions Enhancing Tumor (ET), Whole Tumor (WT), and Tumor Core (TC).	BraTS 2020 Validation Set: Multimodal MRI dataset used for benchmarking brain tumor segmentation algorithms. BraTS 2019 Online Validation Set: Used for further evaluation and comparison	Metric: Dice Score	-BraTS 2020 ET : 78.19% WT : 90.10% TC : 83.98% -BraTS 2019 ET : 77.31% WT : 89.64% TC : 82.55%	V-Net is used in 3D segmentation tasks. Through skip connections It captures volumetric data which make it a strong foundation in 3D tasks [60].	V-Net computational is limited, 3D volumes increase computational cost which makes it challenging to process large volumes.

C. RQ3: What are the Frequently Used Datasets in the Field of Medical Image Segmentation Based on DL?

In image segmentation, datasets are essential as standard references for comparing new segmentation techniques. They act as benchmarks to evaluate the innovation efficiency of a given method in the domain. Table 2 lists the most employed datasets in image segmentation based on DL. These datasets are extensively cited and widely adopted in literature for developing and evaluating DL models. Our selection was guided by three principal criteria: 1) Public Availability and Reproducibility: publicly accessible, well-documented datasets with established use in prior benchmarking studies [63]. This ensures results remain comparable and verifiable by the research community. 2) Annotation Quality: Each dataset provides high-quality expert annotations while capturing inherent challenges of medical imaging including small lesion detection and significant class imbalance across multiple modalities [64]. 3) Generalizability: While many existing studies concentrate on frequently occurring conditions, STS extend evaluation to rare disease types. This addresses a critical gap, as datasets restricted to narrow clinical scenarios risk of overfitting [65].

TABLE II. A REVIEW OF WIDELY USED DATASETS IN MEDICAL IMAGE SEGMENTATION

Dataset	Image type	Limits	Description
ISIC [51]	Dermoscopic	-Image Size Variability: The resizing process causes loss of features which affects the segmentation accuracy. - Variation in Image Quality: This variation impacts the model generalization. - Imbalance Classes: Some classes appear more than others which lead to biased model predictions	- International Skin Imaging Collaboration Archive is a prominent open-source resource that provides a vast collection of annotated skin images specifically for skin cancer research and the development of AI algorithms. - It includes over 156,000 images and supports various initiatives aimed at improving skin cancer diagnosis through standardized imaging practices.
LUNA16 [52]	CT	- Limited Annotations: It provides annotations for nodules, which creates an imbalance between positive and negative nodules. - Variability in Image Quality: It includes scans from different laboratories, each having their scanning protocols and resolutions which introduce noise. - Missing Contextual Information: like patient symptoms or medical records which are important in diagnosis.	The LUNA16 dataset is split into two distinct sections. The first is designated for training while the second is for testing. The training set undergoes a comprehensive data augmentation process aimed at increasing the diversity and robustness of the dataset. This augmentation is crucial in order to enhance performance of machine learning algorithms by providing them with varied examples to learn from.
BUSI [49]	Ultrasound images	- Limited Number of Images: it is relatively small compared to other datasets. - 2D Images: The dataset contains only 2D ultrasound images. - Longitudinal Data: Missing information about the progression of the lesions over time.	It's composed of grayscale images, comprising 780 graphics sourced from 600 individuals. All images are in PNG extension and resold ion of 500 × 500 pixels. The dataset is classified into three distinct classes: Normal, Benin and Malignant.
STS [50]	PET/CT	- Limited patient data available due to the rareness of diseases. - Missing Data: incomplete data for complex variables. - Imbalance Data: Some diseases have few instances in a dataset.	Soft Tissue Sarcoma is a dataset consisting of 51 scans. Regions containing tumours are delineated by a radiologist, and this information was subsequently transferred to the corresponding FDG-PET images utilizing MIM as software. Additionally, for edema cases, data related to hidden edema has been incorporated into the dataset.

D. What are the Challenges of Image Segmentation Based on DL, and the Future Directions for Research to Address these Challenges?

This section delves into the challenges researchers encounter in addressing the complex array of unresolved issues. It discusses obstacles like data limitations and technological constraints, emphasizing the need for innovative solutions. Looking ahead, we outline potential avenues for future research.

1) Challenges: This task is inherently complex due to the intricacies of human anatomy, the vast variability in disease presentations, and the subtle distinctions between normal and abnormal tissues. Technically, these challenges are further compounded by the limitations of imaging modalities, where factors such as noise, resolution, and contrast can significantly impact image quality. While deep learning offers a promising approach for automated segmentation, several data-related obstacles must be addressed, including the need for large and annotated datasets.

a) Data diversity and quality: Data diversity and quality encompass the range and accuracy of the datasets employed in training and assessing models. Ensuring that these datasets capture the real-world variability requires diverse and representative patient demographics and disease conditions. A major challenge lies in the typically small and limited datasets available, especially for rare conditions, as medical datasets often lack sufficient patient numbers and diversity. Q. Zhang et al., [54] identified a key limitation in bladder tumor segmentation using DL techniques, emphasizing the challenge encountered by the restricted size and diversity of available datasets. Luo et al. [50] emphasized the challenge of this inconsistency, particularly in the context of bladder tumors. They pointed out the noticeable morphological differences in tumors across various grades and stages, which manifest in cystoscopic images through variations in lesion appearance, color, and size. This inherent variability underscores the difficulty in diagnosing and treating bladder cancer, highlighting the need for models capable of handling a wide array of tumor characteristics. Medical images frequently exhibit variations in intensity and are prone to noise, often resulting from variations in imaging devices, configurations, or patient alignment. These discrepancies can interfere with a model's ability to learn essential features necessary for precise segmentation. Labeling medical images demands specialized expertise, making it both time-intensive and costly. The lack of labeled data limits the availability of high-quality training resources necessary for developing robust DL models.

b) Validation and generalizability of the model: The goal is to confirm that DL models achieve high accuracy in controlled environments and maintain consistency and performance across a variety of real-world clinical scenarios. This involves ensuring models need to be able to generalize across various datasets, with a strong focus on the necessity of thorough clinical validation. Generalizability is critical to ensure that models stay adaptable and aren't overly dependent on the specific traits of data. Q. Zhang et al. [54] discussed the important issues of generalizability and external validation. Real-world clinical validation emphasizes the indispensability of testing DL models in genuine clinical settings, rather than relying solely on simulation-based evaluations. It underscores the importance of proving the models' effectiveness and accuracy in actual medical practice. This validation is vital to confirm models' assistance in diagnosis and treatment planning.

c) Enhancing computational performance and optimizing system architecture: Model-driven approach learns and refines uncertain regions between tumors and tissues, which lead to better accuracy of tumor boundary. Regarding medical image segmentation with deep learning, computational efficiency and architecture optimization are centered on improving model performance while minimizing computational requirements. Improving the collaboration among various components of a deep learning model is crucial for improving its overall efficiency. This optimization ensures that the model's layers and operations collaborate smoothly, resulting in quicker processing without sacrificing accuracy. This is especially important in medical imaging, where models must process intricate data with high efficiency. Luo et al [50] examines the challenges related to efficiency in contextual coordination. To overcome these challenges, they proposed a context coordination module (CCM) that integrates multi-scale contextual information with axial attention. This technique improves the model's ability to focus on tumor regions while optimizing the analysis of feature maps, significantly advancing the efficient management of contextual information for complex segmentation tasks. Furthermore, optimizing model parameters and computational time involves fine- tuning the model to achieve the ideal balance among computational efficiency and segmentation accuracy. Such optimization is essential for developing models that perform efficiently on available hardware, particularly in clinical settings where time sensitivity is paramount.

2) Future research directions

a) Re-configuration: Segmentation algorithms differ depending on the imaging modalities, such as CT, MRI, and X-ray, as well as the specific tissues or anomalies under investigation. Reconfigurable computing allows for the customization of hardware to better accommodate the varying needs of these algorithms ensuring optimal performance. Overall, reconfigurable computing is crucial for medical image segmentation [53], especially when combined with deep learning techniques. It delivers efficient energy-conserving, customizable, and parallel processing capabilities, which are vital for ensuring accurate medical image analysis, ultimately improving patient outcomes through faster and more reliable diagnostics.

b) Neuromorphic learning: Neuromorphic learning, utilizing neural models in both hardware and software, holds immense potential in medicine, particularly for processing complex medical images more efficiently than traditional algorithms. These systems mimic the human brain's structure and parallel processing capabilities, offering advantages like low power consumption and high computational capacity.

However, obstacles like model complexity, limited understanding, and the absence of standardization restrict their application in medical image segmentation.

V. Conclusion

This review seeks to present an overview of studies focused on DL applications in medical image processing and analysis. Through a systematic approach, we have curated an extensive collection of contributions, categorizing them into various segments. Our primary objective is to address critical research questions regarding the evolving intersections of DL and medical imaging, highlighting emerging trends and advancements in the field. We investigate the most widely used DL algorithms, emphasizing their advantages and limitations. This paper delves into the challenges surrounding DL-based medical image segmentation and proposes future research directions to overcome these hurdles. As a comprehensive review, it serves as a key resource, offering critical insights to drive advancements in the field and benefit both researchers and stakeholders.

Acknowledgment

The authors wish to acknowledge the support provided for this research by: Author contribution. Amine Berquedich: Data curation, Writing - Original draft preparation: Review article; Ahmed Zellou Supervision, proofreading

References

[1] B. Azad, R. Azad, S. Eskandari, A. Bozorgpour, A. Kazerouni, I. Rekik, and D. Merhof, "Foundational models in medical imaging: A comprehensive survey and future vision," *arXiv preprint* arXiv:2310.18689, 2023, doi: https://arxiv.org/abs/2310.18689

[2] T. Shiang, E. Thai, R. Hanson, R. Lacson, S. Saini, C. S. Giess, D. W. Bates, and R. Khorasani, "Inpatient imaging utilization and radiology workload: Trends of the past decade and through the COVID-19 pandemic," *Journal of the American College of Radiology (JACR)*, vol. 22, no. 9, pp. 979–989, 2025, doi: https://doi.org/10.1016/j.jacr.2025.05.011

[3] L. H. Portnow, L. Choridah, K. Kardinah, T. Handarini, R. Pijnappel, A. M. J. Bluekens, L. E. M. Duijm, P. K. Schoub, P. S. Smilg, L. Malek, J. W. T. Leung, and S. Raza, "International interobserver variability of breast density assessment," Journal of the American College of Radiology (JACR), vol. 20, no. 7, pp. 671–684, 2023, doi: https://doi.org/10.1016/j.jacr.2023.03.010

[4] B. Sistaninejhad, H. Rasi, and P. Nayeri, "A review paper about deep learning for medical image analysis," Computational and Mathematical Methods in Medicine, vol. 2023, Article ID 7091301, 2023, doi: https://doi.org/10.1155/2023/7091301

[5] I. D. Mienye and T. G. Swart, "A comprehensive review of deep learning: Architectures, recent advances, and applications," Information, vol. 15, no. 12, p. 755, 2024, doi: https://doi.org/10.3390/info15120755

[6] K. Petersen, R. Feldt, S. Mujtaba, and M. Mattsson, "Systematic Mapping Studies in Software Engineering," Jun. 2008, doi: https://doi.org/10.14236/ewic/ease2008.8

[7] D. Weyns, P. Avegriou, R. Calinescu, S. M. Hezavehi, R. Mirandola, and D. Perez-Palacin, "Specification architectural viewpoint for benefit-cost-risk-aware decision-making in self-adaptive systems," arXiv preprint arXiv:2211.17218, 2022, doi: https://arxiv.org/abs/2211.17218

[8] K. Ahmad, M. Abdelrazek, C. Arora, M. Bano, and J. Grundy, "Requirements engineering for artificial intelligence systems: A systematic mapping study," arXiv preprint arXiv:2212.10693, 2022, doi: https://arxiv.org/abs/2212.10693

[9] M. H. Tanzil, S. Chowdhury, S. Modaberi, G. Uddin, and H. Hemmati, "A systematic mapping study of crowd knowledge enhanced software engineering research using Stack Overflow," Journal of Systems and Software, vol. 226, 2025, 112405, doi: https://doi.org/10.1016/j.jss.2025.112405

[10] 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, 106366, doi: https://doi.org/10.1016/j.infsof.2020.106366

[11] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering," Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering - EASE '14, no. 38, pp. 1–10, 2014, doi: https://doi.org/10.1145/2601248.2601268

[12] P. K. Mall et al., "A comprehensive review of deep neural networks for medical image processing: Recent developments and future opportunities," Healthcare Analytics, vol. 4, p. 100216, Dec. 2023, doi: https://doi.org/10.1016/j.health.2023.100216

[13] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, and M. J. Deen, "Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives," Neurocomputing, vol. 444, pp. 92–110, Jul. 2021, doi: https://doi.org/10.1016/j.neucom.2020.04.157

[14] B. Pandey, D. Kumar Pandey, B. Pratap Mishra, and W. Rhmann, "A comprehensive survey of deep learning in the field of medical imaging and medical natural language processing: Challenges and research directions," Journal of King Saud University - Computer and Information Sciences, Jan. 2021, doi: https://doi.org/10.1016/j.jksuci.2021.01.007

[15] L. Ajallouda, F. Z. Fagroud, A. Zellou, and E. B. Lahmar, "KP-USE: An Unsupervised Approach for Key-Phrases Extraction from Documents," International Journal of Advanced Computer Science and Applications, vol. 13, no. 4, 2022, doi: https://doi.org/10.14569/ijacsa.2022.0130433

[16] M. Puttagunta and S. Ravi, "Medical image analysis based on deep learning approach," Multimedia Tools and Applications, Apr. 2021, doi: https://doi.org/10.1007/s11042-021-10707-4

[17] S. Basheera and M. S. Sai Ram, "Convolution neural network–based Alzheimer's disease classification using hybrid enhanced independent component analysis based segmented gray matter of T2 weighted magnetic resonance imaging with clinical valuation," Alzheimer's & Dementia: Translational Research & Clinical Interventions, vol. 5, no. 1, pp. 974–986, Jan. 2019, doi: https://doi.org/10.1016/j.trci.2019.10.001

[18] S.-H. Wang, K. Muhammad, J. Hong, A. K. Sangaiah, and Y.-D. Zhang, "Alcoholism identification via convolutional neural network based on parametric ReLU, dropout, and batch normalization," Neural Computing and Applications, vol. 32, no. 3, pp. 665–680, Dec. 2018, doi: https://doi.org/10.1007/s00521-018-3924-0

[19] C. Tian, B. Zhang, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin, "Deep Learning on Image Denoising: An overview," arXiv (Cornell University), Dec. 2019, doi: https://doi.org/10.48550/arxiv.1912.13171

[20] G. Sood, "A Hybrid Approach of Semantic Weight Based Re-Propagation For Convolutional Neural Networks in Content Based Medical Image Retrieval," pp. 1–7, Apr. 2023, doi: https://doi.org/10.1109/icdcece57866.2023.10151427

[21] D. Bhanu Mahesh, G. Satyanarayana Murty, and D. Rajya Lakshmi, "Optimized Local Weber and Gradient Pattern-based medical image retrieval and optimized Convolutional Neural Network-based classification," Biomedical Signal Processing and Control, vol. 70, p. 102971, Sep. 2021, doi: https://doi.org/10.1016/j.bspc.2021.102971

[22] Wang Xiangcan, J. Wang, and B. Li, "Low dose CT image denoising method based on improved generative adversarial network," Jul. 2022, doi: https://doi.org/10.1109/cacre54574.2022.9834193

[23] V. S. Kumar and V. Jayalakshmi, "Reconstructing the Medical Image by Autoencoder with Stochastic Processing in Neural Network," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 1521–1526, Sep. 2021, doi: https://doi.org/10.1109/icirca51532.2021.9545041

[24] C. Wang, X. Lv, M. Shao, Y. Qian, and Y. Zhang, "A novel fuzzy hierarchical fusion attention convolution neural network for medical image super-resolution reconstruction," Information Sciences, vol. 622, pp. 424–436, Apr. 2023, doi: https://doi.org/10.1016/j.ins.2022.11.140

[25] F. E. Fernandes and G. G. Yen, "Pruning of generative adversarial neural networks for medical imaging diagnostics with evolution strategy,"

Information Sciences, vol. 558, pp. 91–102, May 2021, doi: https://doi.org/10.1016/j.ins.2020.12.086

[26] S. Patel, "An Overview and Application of Deep Convolutional Neural Networks for Medical Image Segmentation," Feb. 2023, doi: https://doi.org/10.1109/icais56108.2023.10073857

[27] S. Devunooru, A. Alsadoon, P. W. C. Chandana, and A. Beg, "Deep learning neural networks for medical image segmentation of brain tumours for diagnosis: a recent review and taxonomy," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 1, pp. 455–483, May 2020, doi: https://doi.org/10.1007/s12652-020-01998-w

[28] L. Xie et al., "Deep label fusion: A generalizable hybrid multi-atlas and deep convolutional neural network for medical image segmentation," Medical Image Analysis, vol. 83, p. 102683, Jan. 2023, doi: https://doi.org/10.1016/j.media.2022.102683

[29] D. Yuvaraj, A. K. Sampath, N Shanmugapriya, Badugu Samatha, S. Arun, and R. Thiyagarajan, "Medical Image Segmentation of Bio-medical Images with Deep Convolution Neural Networks using Ensemble Approach," 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), pp. 1041–1049, Oct. 2022, doi: https://doi.org/10.1109/icosec54921.2022.9952142

[30] Zakaria Rguibi, AbdelMajid Hajami, and Zitouni Dya, "Explaining Deep Neural Networks in medical imaging context," pp. 1–2, Nov. 2021, doi: https://doi.org/10.1109/aiccsa53542.2021.9686919

[31] T. Zhou, Q. Cheng, H. Lu, Q. Li, X. Zhang, and S. Qiu, "Deep learning methods for medical image fusion: A review," Computers in Biology and Medicine, vol. 160, p. 106959, Jun. 2023, doi: https://doi.org/10.1016/j.compbiomed.2023.106959

[32] D. Vasanthi, U. Palani, M. Vishnupriya, and K. Saundariya, "Detection of Tumor by Fusing Medical Images using Artificial Neural Network," 2021 International Conference on System, Computation, Automation and Networking (ICSCAN), pp. 1–6, Jul. 2021, doi: https://doi.org/10.1109/icscan53069.2021.9526521

[33] Z. Chao, X. Duan, S. Jia, X. Guo, H. Liu, and F. Jia, "Medical image fusion via discrete stationary wavelet transform and an enhanced radial basis function neural network," Applied Soft Computing, vol. 118, p. 108542, Mar. 2022, doi: https://doi.org/10.1016/j.asoc.2022.108542

[34] N. Shyamala and S. Geetha, "Fusion model of Modified Wavelet Transform and Neural Network for medical image compression," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), pp. 1676–1682, Aug. 2022, doi: https://doi.org/10.1109/icesc54411.2022.9885564

[35] W. Xu, Y.-L. Fu, H. Xu, and K. K. L. Wong, "Medical image fusion using enhanced cross-visual cortex model based on artificial selection and impulse-coupled neural network," Computer Methods and Programs in Biomedicine, vol. 229, pp. 107304–107304, Dec. 2022, doi: https://doi.org/10.1016/j.cmpb.2022.107304

[36] L. Zhang, M. Pereañez, S. K. Piechnik, S. Neubauer, S. E. Petersen, and A. F. Frangi, "Image Quality Assessment for Population Cardiac Magnetic Resonance Imaging," Advances in computer vision and pattern recognition, pp. 299–321, Jan. 2019, doi: https://doi.org/10.1007/978-3-030-13969-8_15

[37] I. Urbaniak and M. W. Wolter, "Quality assessment of compressed and resized medical images based on pattern recognition using a convolutional neural network," Communications in Nonlinear Science and Numerical Simulation, vol. 95, pp. 105582–105582, Apr. 2021, doi: https://doi.org/10.1016/j.cnsns.2020.105582

[38] O. H. Nagoor, J. Whittle, J. Deng, B. Mora, and M. W. Jones, "MedZip: 3D Medical Images Lossless Compressor Using Recurrent Neural Network (LSTM)," 2022 26th International Conference on Pattern Recognition (ICPR), vol. 4, pp. 2874–2881, Jan. 2021, doi: https://doi.org/10.1109/icpr48806.2021.9413341

[39] H. R. Boveiri, R. Khayami, R. Javidan, and A. Mehdizadeh, "Medical image registration using deep neural networks: A comprehensive review," Computers & Electrical Engineering, vol. 87, p. 106767, Oct. 2020, doi: https://doi.org/10.1016/j.compeleceng.2020.106767

[40] S. Abbasi et al., "Medical image registration using unsupervised deep neural network: A scoping literature review," Biomedical Signal Processing and Control, vol. 73, p. 103444, Mar. 2022, doi: https://doi.org/10.1016/j.bspc.2021.103444

[41] Z. Zheng, W. Cao, Z. He, and Y. Luo, "Progressive anatomically constrained deep neural network for 3D deformable medical image registration," Neurocomputing, vol. 465, pp. 417–427, Nov. 2021, doi: https://doi.org/10.1016/j.neucom.2021.08.097

[42] D. Yang, T. Xiong, and D. Xu, "Automatic Vertebra Labeling in Large-Scale Medical Images Using Deep Image-to-Image Network with Message Passing and Sparsity Regularization," Advances in computer vision and pattern recognition, pp. 179–197, Jan. 2019, doi: https://doi.org/10.1007/978-3-030-13969-8_9

[43] Abirami R and Malathy C, "Detecting Singular Value Decomposition-based WatermarkEmbedded Medical Images using ResNet Neural Network," 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), pp. 1–8, May 2023, doi: https://doi.org/10.1109/accai58221.2023.10201092

[44] J. Ru et al., "Attention guided neural ODE network for breast tumor segmentation in medical images," Computers in Biology and Medicine, vol. 159, pp. 106884–106884, Jun. 2023, doi: https://doi.org/10.1016/j.compbiomed.2023.106884

[45] M. Gao, Z. Xu, and D. J. Mollura, "Interstitial Lung Diseases via Deep Convolutional Neural Networks: Segmentation Label Propagation, Unordered Pooling and Cross-Dataset Learning," Advances in Computer Vision and Pattern Recognition, pp. 97–111, 2017, doi: https://doi.org/10.1007/978-3-319-42999-1_7

[46] A. Veeramuthu, S. Meenakshi, and K. Ashok Kumar, "A neural network based deep learning approach for efficient segmentation of brain tumor medical image data," Journal of Intelligent & Fuzzy Systems, vol. 36, no. 5, pp. 4227–4234, May 2019, doi: https://doi.org/10.3233/jifs-169980

[47] R. T. Subhalakshmi, S. Appavu Alias Balamurugan, and S. Sasikala, "Automatic Segmentation and Classification of COVID-19 CT Image Using Deep Learning and Multi-Scale Recurrent Neural Network Based Classifier," Journal of Medical Imaging and Health Informatics, vol. 11, no. 10, pp. 2618–2625, Oct. 2021, doi: https://doi.org/10.1166/jmihi.2021.3850

[48] M. Ghafoori et al., "Predicting survival of Iranian COVID-19 patients infected by various variants including omicron from CT Scan images and clinical data using deep neural networks," Heliyon, vol. 9, no. 11, pp. e21965–e21965, Nov. 2023, doi: https://doi.org/10.1016/j.heliyon.2023.e21965

[49] B. Shareef, A. Vakanski, P. E. Freer, and M. Xian, "ESTAN: Enhanced Small Tumor-Aware Network for Breast Ultrasound Image Segmentation," Healthcare, vol. 10, no. 11, p. 2262, Nov. 2022, doi: https://doi.org/10.3390/healthcare10112262

[50] S. Luo, H. Jiang, and M. Wang, "C2BA-UNet: A context-coordination multi-atlas boundary-aware UNet-like method for PET/CT images based tumor segmentation," Computerized Medical Imaging and Graphics, vol. 103, p. 102159, Jan. 2023, doi: https://doi.org/10.1016/j.compmedimag.2022.102159

[51] R. Mohakud and R. Dash, "Skin cancer image segmentation utilizing a novel EN-GWO based hyper-parameter optimized FCEDN," Journal of King Saud University - Computer and Information Sciences, Jan. 2022, doi: https://doi.org/10.1016/j.jksuci.2021.12.018

[52] D. Bhattacharyya, N. Thirupathi Rao, E. S. N. Joshua, and Y.-C. Hu, "A bi-directional deep learning architecture for lung nodule semantic segmentation," The Visual Computer, Sep. 2022, doi: https://doi.org/10.1007/s00371-022-02657-1

[53] H. Messaoudi, A. Belaid, D. B. Salem, and P.-H. Conze, "Cross-dimensional transfer learning in medical image segmentation with deep learning," Medical Image Analysis, vol. 88, p. 102868, Aug. 2023, doi: https://doi.org/10.1016/j.media.2023.102868

[54] Q. Zhang, Y. Liang, Y. Zhang, Z. Tao, R. Li, and H. Bi, "A comparative study of attention mechanism based deep learning methods for bladder tumor segmentation," International Journal of Medical Informatics, vol. 171, p. 104984, Mar. 2023, doi: https://doi.org/10.1016/j.ijmedinf.2023.104984

[55] J. Sun, Y. Peng, Y. Guo, and D. Li, "Segmentation of the multimodal brain tumor image used the multi-pathway architecture method based on 3D FCN," Neurocomputing, vol. 423, pp. 34–45, Jan. 2021, doi: https://doi.org/10.1016/j.neucom.2020.10.031

[56] S. Kumar, A. Negi, J. A. Singh, and H. Verma, "A Deep Learning for Brain Tumor MRI Images Semantic Segmentation Using FCN," Dec. 2018, doi: https://doi.org/10.1109/ccaa.2018.8777675

[57] S. Tripathi, R. Wadhwani, A. Rasool, and A. Jadhav, "Comparison Analysis of Deep Learning Models In Medical Image Segmentation," 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 468–471, Jan. 2023, doi: https://doi.org/10.1109/confluence56041.2023.10048822

[58] S. Subramaniam, K. B. Jayanthi, C. Rajasekaran, and R. Kuchelar, "Deep Learning Architectures for Medical Image Segmentation," IEEE Xplore, Jul. 01, 2020. https://ieeexplore.ieee.org/abstract/document/9182808/

[59] J. Wang and X. Liu, "Medical image recognition and segmentation of pathological slices of gastric cancer based on Deeplab v3+ neural network," Computer Methods and Programs in Biomedicine, vol. 207, p. 106210, Aug. 2021, doi: https://doi.org/10.1016/j.cmpb.2021.106210

[60] Z. Wang, Y. Zou, H. Chen, P. X. Liu, et J. Chen, "Multi-scale features and attention guided for brain tumor segmentation", Journal Of Visual Communication And Image Representation, vol. 100, p. 104141, 2024.

[61] Bindu Neelam, Pavan Kumar Palakayala, Kusiyo Mbangweta, Karthik Raparla, and S. A. Devi, "FCN Based Deep Learning Architecture for Medical Image Segmentation," vol. abs 2104 14353, pp. 556–562, Jul. 2023, doi: https://doi.org/10.1109/icecaa58104.2023.10212108

[62] S. M. Mäenpää and M. Korja, "Diagnostic test accuracy of externally validated convolutional neural network (CNN) artificial intelligence (AI) models for emergency head CT scans – A systematic review," International Journal of Medical Informatics, vol. 189, 2024, 105523, doi: https://doi.org/10.1016/j.ijmedinf.2024.105523.

[63] C.-Y. Lin, S.-M. Guo, J.-J. J. Lien, W.-T. Lin, Y.-S. Liu, C.-H. Lai, I.-L. Hsu, C.-C. Chang, and Y.-L. Tseng, "Combined model integrating deep learning, radiomics, and clinical data to classify lung nodules at chest CT," La Radiologia Medica, vol. 129, no. 1, pp. 56–69, 2024, doi: https://doi.org/10.1007/s11547-023-01730-6.

[64] L. Cui, M. Xu, C. Liu, et al., "Towards reliable healthcare imaging: A multifaceted approach in class imbalance handling for medical image segmentation," Interdisciplinary Sciences: Computational Life Sciences, vol. 17, pp. 614–633, 2025, doi: https://doi.org/10.1007/s12539-025-00726-2.

[65] N. Kiryati and Y. Landau, "Dataset growth in medical image analysis research," Journal of Imaging, vol. 7, no. 8, p. 155, 2021, doi: https://doi.org/10.3390/jimaging7080155.

[66] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," Nature Methods, vol. 18, no. 2, pp. 203–211, 2021, doi: https://doi.org/10.1038/s41592-020-01008-z.

[67] M. Antonelli, A. Reinke, S. Bakas, et al., "The Medical Segmentation Decathlon," Nature Communications, vol. 13, no. 1, p. 4128, 2022. DOI: 10.1038/s41467-022-30695-9

[68] S. G. Finlayson, A. Subbaswamy, K. Singh, et al., "The clinician and dataset shift in artificial intelligence," New England Journal of Medicine, vol. 385, no. 3, pp. 283–286, 2021. DOI: 10.1056/NEJMc2104626

[69] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: Methodological failures and recommendations for the future," NPJ Digital Medicine, vol. 5, no. 1, p. 48, 2022. DOI: 10.1038/s41746-022-00592-y

[70] A. Subbaswamy and S. Saria, "From development to deployment: Dataset shift, causality, and shift-stable models in health AI," Biostatistics, vol. 21, no. 2, pp. 345–352, 2020. DOI: 10.1093/biostatistics/kxz041

[71] E. Wu, K. Wu, R. Daneshjou, et al., "How medical AI devices are evaluated: Limitations and recommendations from an analysis of FDA approvals," Nature Medicine, vol. 27, no. 4, pp. 582–584, 2021. DOI: 10.1038/s41591-021-01312-x

Modeling and Analyzing Malware Behavior in Virtual Networks Using EVE-NG

Maria-Mădălina Andronache[1], Alexandru Vulpe[2], Corneliu Burileanu[3]
Research Institute "CAMPUS", National University of Science and Technology Politehnica Bucharest, Bucharest, Romania[1]
Telecommunication Department, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania[2]
Speech and Dialogue Research Lab, National University of Science and Technology Politehnica Bucharest, Bucharest, Romania[3]

***Abstract*—Malicious attacks have become increasingly common in all organizations and systems. The continued evolution of such software aims to extract information from diverse systems. Therefore, the objective of this study is to introduce another approach to analyze some network attacks, within a virtual infrastructure, through multi-vendor network emulation software (Emulated Virtual Environment–Next Generation - EVE-NG). Basically, through emulated resources, the aim is to implement a complex network, which also includes a Security Information and Event Management System (SIEM), which can detect some attacks, both from the network area (carried out by malicious attackers) and through malicious files (from the public resources area), that are accidentally or intentionally downloaded by certain users. Within this environment, various scenarios can be implemented to simulate the real production environment, in order to test network vulnerabilities, but also to improve some methods for learning network attack and defense modes. In the experiments performed, the SIEM system detected most of the simulated attacks, but failed to distinguish between the displayed alarms so that the alerts could indicate the type of attack. Thus, the potential of EVE-NG for simulating and analyzing the behavior of malware is demonstrated.**

***Keywords*—EVE-NG; network attacks; network defense; SIEM; network architecture**

I. Introduction

In the current security landscape, the primary actors are continuously evolving, shaped by intensifying threats and by the new and more effective defense against the aforementioned threats. As a current way of working, in both camps, new methods based on artificial intelligence and automation of various attacks or, on the contrary, new, much more effective, defense methods, are emerging. In practice, advances are being made both in attack methods, as well as in the area of zero-trust architecture and security policies based on identification methods. In addition, key aspects are also becoming the use, on an increasingly large scale, of cloud resources and IoT technologies, which, paradoxically, increases security risks and zero-day vulnerabilities.

As networks evolve, network devices become more secure and more capable of advanced access control or observability policies with each update. In practice, most network device providers also integrate artificial intelligence or machine-learning techniques to capture various logs and detect anomalies, both in the device's own operation and at the level of a party directly involved with the device. Thus, through them, a better segmentation and a more efficient combination of network capabilities with security policies can be achieved. If a few years ago, the emphasis was on the hardware area, currently, the most intensively researched part is the software.

According to study [1], understanding how cyberattacks work and their key terminology is essential in order to know how different types of threats work and which are the most important methods to respond to some attacks. Therefore, from the attack name, one can infer commonly used strategies, frequently adopted by that malicious file.

From study [2], the key aspects that define effective cybersecurity measures are:

- Providing information that includes key characteristics about the threat, quickly.
- The ability to provide details about the main risks of the victim organization.
- Must contain clear information about the impact, which prevents the formation of a false sense of security at the organizational level.
- Requires clear and concrete information about the attack.
- Include recommended actions to prevent damage and the expansion of the attack.

As networks become larger and more interconnected, understanding how different types of attacks work and how defenses respond has become a critical area of study in cybersecurity. Experimental analysis is essential for this purpose, but conducting such research on production systems can be both expensive and dangerous, potentially exposing real networks to damage. As a result, the use of virtualized environments to simulate attacks and evaluate defenses has attracted significant attention from both researchers and educators.

Emulated Virtual Environment-Next Generation (EVE-NG) [3] is a working environment used to build various virtual network architectures, which can include equipment from several vendors. Its purpose is to learn numerous concepts and working methods, in an easier way. For an attack-type testing environment on a real network topology, as is the case in this paper, EVE-NG is an ideal environment because it facilitates access to virtualized resources while keeping the production environment intact. Thus, it works similar to a sandbox, only replicating the entire network, not just a specific part of it or a specific workstation.

EVE-NG is not formally intended for executing malicious software or simulating attack payloads. Despite this limitation, its isolated design and versatile networking features make it a practical option for controlled cybersecurity experimentation. In this paper, is explored how EVE-NG can be adapted to perform secure simulations of various network-level attack scenarios while maintaining system integrity and ethical boundaries.

Considering the information presented in study [4], emulators such as EVE-NG offer a virtualized solution closer to real working environments, by including, in the test environment, the real code images of the devices that are emulated. Through this way of working, a more accurate assessment of the impact on workstations or network devices can be achieved, by testing some features much closer to the real environment. Thus, EVE-NG is an extremely useful virtualized environment for learning the characteristics of real equipment, without disturbing their functioning in the production environment.

Malicious software is a piece of code or an entire program designed to exploit various network resources. Its main purpose is to steal data, obtain material benefits (by ransoming resources) or destroy information. The types of malicious files are diverse, but the most commonly used for normal users are viruses, adware or spyware. Within an organization's resources, the most used malicious files are ransomware or viruses, worms and trojans because the information that is wanted to be extracted is much more valuable, from a financial gain point of view, and, thus, the attack must be more sophisticated. Also, considering the latest trends in tactics, the newest ones used include combinations of multiple malware categories, changes in its own code, to make it easier to replicate, or the use of legitimate resources, already present on the victim system (living-off-the-land).

Considering the aspects presented in the paper [5], the impact of malicious software is different, depending on their intention. Thus, while some are directed to a station or a system, others are used to be spread to a part of the network or the entire network, to infect a larger number of devices. In addition, certain types of malicious files (worms, trojans, ransomware) have higher levels of aggression, even leading to the permanent interruption of some services, while adware or keyloggers collect information about the device for commercial purposes or for the purpose of information theft. In order to be characterized efficiently and completely, a certain type of malware must be analyzed in a hybrid way, both statically (without being executed) and dynamically (through its execution).

External network-level attacks complement malicious file-based threats. Therefore, the usual modus operandi is that external malicious actors can perform DoS (Denial of Service) attacks to overload services, phishing attacks to trick users into running malicious software, or brute force attacks to gain initial access. After this step, introducing a malicious file into the network is an easy task. If there are also malicious actors within the internal area, various privilege escalations, data exfiltration or even the intentional downloading of malicious files can be used. The impact comes from changes to the configurations on network equipment, deactivating access control policies, account blocking or lateral movements. Citing [6], external attacks are carried out by unauthorized users, who enter the network in the form of intrusion. Their specific modus operandi includes launching and intercepting malicious packets to gain access to private network resources. On the other hand, internal attacks are launched directly from the private area and include violating internal rules, with the aim of degrading network performance. It should be noted that the threat actor pool and the number of security incidents is growing, which means that, although it may seem counterintuitive, organizations with fewer devices are increasingly targeted. Thus, it is imperative to develop effective defense policies and understand, in as much detail as possible, how network resources and attacks work.

The experiments conducted in this study focus on replicating and analyzing specific attack behaviors to observe how threats affect network traffic and device performance. The proposed approach demonstrates that even within the confines of a non-dedicated simulation tool, significant insights into attack mechanisms and defense responses can be obtained. Furthermore, this framework provides a cost-effective and reproducible environment for cybersecurity research, giving students and professionals the opportunity to study real-world threats without jeopardizing live systems. Also, the results encourage further exploration of hybrid simulation methods that combine traditional network tools with security-focused experimentation, helping to bridge the gap between theoretical knowledge and practical defensive practice.

Despite significant progress in malware analysis techniques, existing approaches remain limited in their ability to capture the full complexity of real-world attack behavior. Current methods also focus primarily on host-level changes, overlooking crucial network-level behaviors such as command-and-control communication or lateral propagation. These limitations restrict the accuracy, scalability, and reproducibility of malware research. Consequently, there is a growing need for adaptable, network-oriented frameworks that can safely simulate attack behaviors and analyze their effects under controlled yet realistic conditions.

Despite the availability of malware sandboxes, there is limited research integrating full-network emulation platforms like EVE-NG with SIEM systems for behavioral malware analysis. So, the main objectives of this study are:

- to emulate a realistic multi-branch enterprise network using EVE-NG,
- to simulate network and malware-based attacks,
- to evaluate SIEM performance in detecting and analyzing intrusions.

This paper is structured as follows. Section II and III review the existing literature and background work in the cybersecurity field. Section IV outlines the methodology employed in developing and implementing the proposed system. Section V presents the results obtained from analyzing malware samples and network attacks. Section VI discusses the implications of these findings, and Section VII concludes by summarizing the key insights and suggesting directions for future research.

II. Related Work

To achieve a comprehensive analysis of how network-level attacks or malware work, it is essential to examine relevant research efforts and commonly used tools in this field. Therefore, this section aims to provide a deeper insight into the current state of research, thereby positioning the present study in the broader cybersecurity landscape.

In study [5], a method for analyzing malicious files was implemented using a virtual test environment and a combination of methods such as dynamic analysis and Windows event logging as data input. To make this possible, a database containing both non-malicious samples (including well-known applications such as YouTube or those from sites such as Softonic) and malicious samples (from the open-source Malware Bazaar database [7]) was used. Following experiments comprising 60% legitimate and 40% malicious files, it was found that an MLP-type machine-learning algorithm achieved a remarkable accuracy of 91.2%. In this paper, malware samples from the same database were used, but the testing methods of these samples are different, using a classic sandbox and processes monitor approach to observe malicious behavior.

Another important work in the literature is [8], where an application of malware detection in systems that include Active Directory, is presented. For this detection, the authors used a new intrusion detection system based on provenance, called HADES. Its advantage, compared to conventional ones, is given by the tracking of the anomaly identified at the level of the entire network, through Active Directory and connection sessions. Basically, HADES works as a system that integrates the specific aspects of a SIEM system, bringing the observability context to the network level, and of an efficient intrusion detector of the Active Directory concept. In the current work, a SIEM system is also used, which leads to the capability of being able to observe intrusions within the network architecture presented in the next section. The difference between the two papers consists in the way of capitalizing on the information brought by the logs from the SIEM system and the attacks used in the experiments.

In research [9], a specific network aspect is presented, which includes the concept of DNS (Domain Name System) spoofing. This aspect includes installing malware on the victim computer and introducing a rogue DHCP (Dynamic Host Configuration Protocol) server into the network. The purpose of this paper includes performing a DNS spoofing and phishing attack. To perform these tests, a DNS spoofing engine from the Python programming language is also used, through which the web interface of a site is modified to replicate a legitimate one, and the phishing attack is performed. The test environment used for these experiments includes the Kali Linux operating system and the multi-vendor network emulation software (EVE-NG). This study uses the same test environment; however, the network architecture and experiments are more complex than in the cited paper. Table I shows comparison of related work.

TABLE I. Comparison of Related Work

Study	Environment	Methodology	Attack Type	Detection Mechanism	Accuracy / Outcome	Key Contribution
[5]	Virtual test environment	Dynamic analysis + Windows event logs	Malicious/benign file samples	MLP-based machine learning	91.2% accuracy	File-level malware classification using hybrid analysis
[8]	Enterprise network with Active Directory	Provenance-based IDS (HADES)	System/network anomalies	Rule tracing via AD sessions	Qualitative results, improved network visibility	Network-level anomaly tracking using provenance data
[9]	EVE-NG full-network emulation	Malware installation + rogue DHCP + DNS spoofing	DNS spoofing and phishing (web interface tampering)	Network and host-level logging	Qualitative demonstration of successful DNS spoofing/phishing	Demonstrated attack chain combining rogue DHCP and DNS spoofing to enable phishing
[10]	EVE-NG full-network emulation	Malware installation + rogue DHCP +	DHCP starvation	Network and host-level logging	Qualitative demonstration of DHCP attacks	Demonstrated DHCP attack
[11]	EVE-NG environment simulating real firewalls	Full-network emulation of production systems	DoS and intrusion attacks	Log and alert analysis from virtualized network devices	Similar setup to present work but focused on device configurations	Modeling of real firewall behavior and testing network attack responses
[12]	EVE-NG vs. Packet Tracer and GNS3	Comparative evaluation of network emulators	General network simulation	Feature and resource analysis	Identified firmware-dependency limitation	Benchmarking of EVE-NG's strengths and constraints
[13]	Broad cybersecurity framework (Wireshark, VirusTotal, Google Rapid Response)	Literature-based analysis of cybersecurity evolution	Incident response, malware detection	Theoretical and empirical synthesis	Comprehensive understanding of cybersecurity methods	Overview of tools and practices for malware discovery and response
This work	**EVE-NG full-network emulation**	**Network and host-level logging, SIEM (OSSIM)**	**Nmap, hping3, real malware samples**	**Rule-based SIEM correlation**	**Binary detection (logged / not logged)**	**Evaluation of SIEM visibility for real attacks in emulated networks**

The aspects presented in the paper [10] include similar concepts to the paper [9] because network-level attacks are also used. In this case, the focus is on the DHCP protocol, and the attack brought to the fore is DHCP starvation. The objective of the paper is to have an efficient way of authenticating clients through the DHCP protocol and to prevent DHCP replay attacks. Thus, through the Python programming language, a token exchange is generated that includes the pre-shared key (PSK), MAC address, PIN and system time of the clients. The testing environment was similar to the one present in this paper, EVE-NG, but the reason for this paper did not include security concerns, but only efficient operation, with minimal consumption of hardware resources.

Considering [11], it is concluded that the common operating systems, Linux and Windows, within a network, are protected quite inefficiently if the protection of network devices is missing. Thus, in this work, through the EVE-NG simulation environment, real network devices (firewalls) are modeled, which constitute a complete attack testing environment. As is also specified in this work, the environment created simulates a real production environment. It is mentioned that the experiments carried out in the cited paper are similar to those present in this one, the key differences between the two being given by the fact that, in this paper, the emphasis is not placed on the equipment configurations, but on the results obtained following attacks. Another essential difference is that, in this work, a SIEM type system is implemented, while in the cited work the network security is done through a firewall type device (Web Application Firewall).

Taking into account paper [12], a comparison of the EVE-NG emulation software with other similar tools such as Cisco Packet Tracer and GNS3 is considered. The authors specify that EVE-NG is a robust environment for complex network scenarios, an aspect also certified within this work and has the potential to be an extremely good test environment for the accumulation of knowledge. However, to the authors' claims, an unfavorable aspect of this environment must also be considered: network resources are unavailable unless the firmware images are pre-obtained. This aspect implies having different user accounts or different licenses, otherwise downloading resources is sometimes impossible.

From study [13], a comprehensive understanding of the field of cybersecurity is provided. Thus, useful information can be extracted from the work regarding how incident management is effective at the moment, its development throughout history or the most effective tools needed for malware detection. Throughout the study, a comparison is also made on the detection of a malicious file, which is approached from the perspective of several traditional methods, in relation to those used in antivirus systems. The public resources presented include tools such as Wireshark [14] (which is also used in the experiments in this work) VirusTotal [15] and Google Rapid Response (GRR) [16], used in malware discovery and analysis. However, an essential idea from this work remains the fact that no hardware or software approach can completely detect all forms of malicious software (neither from the area of traditional methods nor from the area of Machine-learning or cloud).

The references cited throughout this section highlight various approaches similar to the one in this paper. Some of them include key concepts from the cybersecurity framework, some include aspects about malicious files, and some include the same testing environment, such as EVE-NG. However, the present paper presents a much more complex network architecture and a way to test the EVE-NG environment different from the aspects in the literature, bringing added value to the field of cybersecurity and to the methods of learning the detection of attacks or software with malicious potential. In contrast to previous studies that examined individual components or specific protocols, this work combines full-network emulation with the execution of actual malware samples and centralized monitoring through a SIEM system.

III. Background Work

Implementing a complex network architecture in a simulated environment emphasizes reproducing production characteristics. Thus, aspects such as border or core routing, network segmentation, access control policies, firewall rules and datacenter area are imminently integrated. In this work, in addition to all of these, elements of malicious attack and network attacks were also included, to model a more unusual behavior of the network and test it under unfavorable conditions. However, EVE-NG is not an environment designed for such operations, but rather for didactic and research purposes.

Another important aspect to mention is the SIEM system. AlienVault OSSIM [17] performs centralized log collection, both from the workstations and from the network devices, their correlation in real time (or in the shortest time available), their display and the creation of alerts, depending on various events or predefined thresholds.

For these things to be possible, an agent is installed at the level of each device in the network, through which the connection with the log manager and its dashboard is ensured. The goal is to identify, analyze and respond, in the most efficient way, to various intrusions or attacks on the network. Platforms such as Wazuh [22], Splunk [23] and QRadar [24] have been studied and analyzed and even compared in terms of their log collection, correlation and alerting capabilities and what makes them different are the characteristics of scalability, detection accuracy and integration flexibility.

Being one of the key subjects of the experiments carried out in this work, the cyberattacks and malicious files require additional explanations. Thus, if we consider network-level attacks, such as network scanning and denial of service, which are more easily implemented in EVE-NG, it must be borne in mind that the purpose of performing them was to test the capabilities of the working environment and create customizable alerts. For the malicious files, the test environment created was the demanding one.

Thus, having the attacks at the network level already performed, some files with malicious software were needed. Although there are a lot of pre-labeled databases, the purpose of these work scenarios was not simply to adapt them, but to perform experiments with real samples. To obtain these samples, the MalwareBazaar Database [7] was used.

This open-source resource includes various types of malicious files specific to Windows and Linux operating systems. The method of choosing the samples did not include a predefined algorithm, but rather the use of a ransomware file and a malware was desired, in order to observe the impact on the work network.

Due to their unpredictable nature, these files were chosen with the aim of carrying out experiments, which can contribute to improving the literature in the field of cybersecurity.

All these aspects, along with other, more advanced details, will be found in the following sections.

IV. METHODOLOGY

Monitoring and analyzing network traffic are essential aspects in protecting infrastructures against cyberattacks. Network anomaly detection aims to identify unusual behaviors such as malware attacks, data theft or unauthorized access attempts. This aspect is imminent within infrastructures that include various types of cybersecurity measures and are essential for any network infrastructure.

The process of detecting them includes certain stages that are based on advanced data analysis or machine learning techniques.

The first one is network data collection: data packet analysis (which includes information about its source and destination, the protocols used and the packet sizes), data flow analysis (NetFlow/sFlow – which can lead to understanding the general behavior of network traffic and can allow the identification of anomalies at the session level) and security log analysis (generated by security equipment, such as firewalls or Intrusion Detection System/Intrusion Prevention System – IDS/IPS).

The second one is creating a "normal" traffic model through identifying traffic typologies (by classifying it according to the nature of various characteristics such as applications and users Web traffic, File Transfer Protocol connections, etc.) and user behavior (by identifying "normal" user behavior).

The third one is continuous monitoring and anomaly detection through machine learning algorithms (identifying anomalies through behavioral models). An example, in this case, is volumetric analysis, which determines Distributed Denial of Service (DDoS) attacks and anomalies in user behavior (deviations from the usual activity of a user or a group of users).

The last one is creating alerts and automatic responses by isolating affected resources and further investigating the attack.

Within the network architecture created for the test scenarios that will be developed throughout the experiments in this work, several different parts of the network were included. Some of these are the data center area (where several types of servers were implemented), the data acquisition and detection part (given by the SIEM Manager), the monitoring area (given by monitoring servers and various end-devices) and some branch areas (that have specific functions). These details can be found in Fig. 1, where some abbreviations appear, whose explanation is necessary.

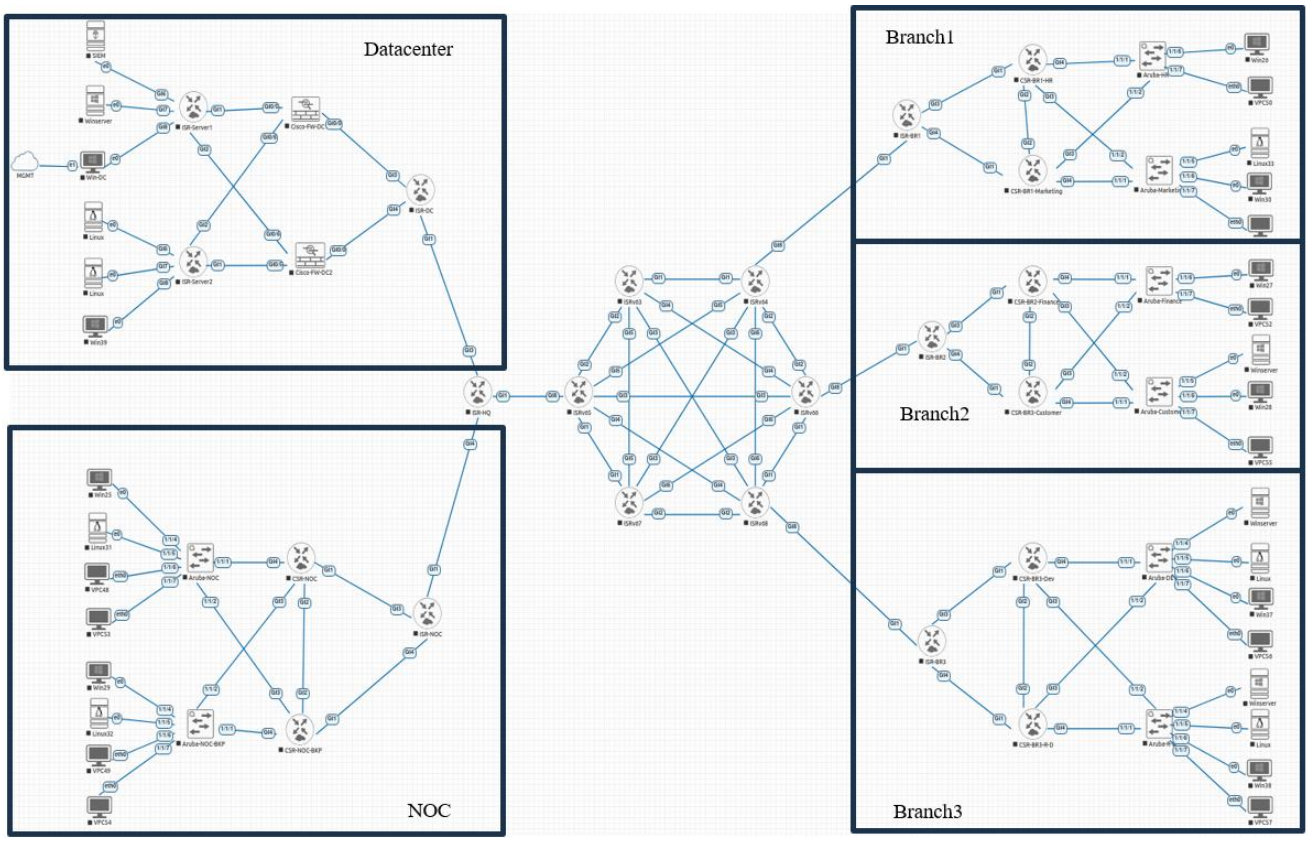


Fig. 1. The analyzed network architecture.

In the datacenter block, located in the upper left part of the figure, there is a SIEM system implemented, the external network management area, which ensures file transfer and other types of communications with the external environment, servers with the Windows Server, Linux or Kali Linux operating system, client stations with Windows and Linux operating systems, ISR routers (to ensure communication from the datacenter area with the rest of the network) in a configuration that also ensures a backup, in the event of an unforeseen event, firewalls (to protect access resources in and from the datacenter) and a router through which data is transmitted to the branch area and to the NOC (network operations center) area.

The NOC block and the branches area (located in the lower left and throughout the right side of the architecture) include Windows and Linux client stations, Aruba access switches, routers in a backup configuration to ensure communication with the external environment and the efficient transmission of CSR monitoring data and ISR gateway routers.

The central block features a full mesh network infrastructure, in which all ISR routers are connected to other routers of the same type. In this way, the internet resources within this infrastructure are simulated. Practically, the test scenarios will include both attacks from the datacenter area, although these are quite rare, and from the NOC (Network Operations Centre) area, but also from the branch area. Therefore, in this way, it will be possible to test the acquisition method and the efficiency of the SIEM system, as well as the propagation method of various types of attacks at the network level.

The entire test infrastructure was implemented using the EVE-NG solution, along with various network components, downloaded from the public resource area. The proposed network architecture included an available hardware environment of approximately 2 TB storage and 512 GB RAM, so that all equipment can function properly, at the same time. Several types of equipment were chosen, from several suppliers, so that the scenario was as realistic as possible and closer to a real work environment. All monitoring infrastructure and agents were deployed as guest virtual machines within the EVE-NG topology.

The SIEM server was provisioned on a virtual machine, also within the EVE-NG platform and configured according to the vendor's documentation.

The branches were created to perform specific functions, such as:

- Human Resources Department – the area that includes the human resources management part.
- Marketing Department – the area that deals with promoting the organization's resources.
- Financial Department – the area that ensures the financial side of the organization.
- Customer Support Department – the area that ensures the interface with the organization's customers.
- Software Development Department – the area that implements IT resources.
- Research Department – the area that innovates IT environments or other types of environments in the organization.

These are key areas of any medium-sized organization and require permanent monitoring and constant communications with the datacenter area.

The NOC part ensures the monitoring of resources, both in the datacenter area and in the branch area, in order to have a centralized view at the level of the entire network. In fact, although the SIEM system has its physical connection in the datacenter area, it is accessed, through its graphical interface, also from the monitoring area. Through this way of working, important resources and attacks at the network level are constantly visualized, with minimal possibilities for the NOC team not to notice an attack.

Agent installation was performed inside each endpoint virtual machine using the native installer for the respective operating system (using, in most cases, the AlienVault documentation).

Agent registration used the SIEM server's EVE-NG internal IP address, TLS keys, and authentication tokens.

Local log forwarding was configured to point to the local agent, which in turn forwarded the logs through an encrypted channel to the SIEM server virtual machine.

Basically, the entire infrastructure worked on the agent-server-dashboard principle, where the agent provided logs to the server, which in turn displayed them and created alerts based on them. Therefore, the logging flow can be summarized as: (emulated host VM) → (local syslog/agent inside the VM) → (encrypted agent channel) → (SIEM server VM inside EVE-NG) → (indexer/correlation engine) → (dashboard).

This aspect leads to the idea that, for a medium business environment, such as the one represented in the figure below, the most expensive equipment will not always be used to standardize the network, but there will also be equipment that includes a lower cost, but with functionalities similar to the "premium" ones.

Attacks at the network level, as well as its actual architecture, provided an ideal framework for observing the behavior of a real infrastructure in the event of an intrusion. Therefore, although it is a simulated environment, through the EVE-NG platform, this framework can constitute a secure basis for performing several types of work scenarios, in a controlled environment, without affecting the production infrastructure.

All malware traffic came from attacker virtual machines created in the EVE-NG topology, so that all malicious activity remained confined to the emulated environment. Widely adopted open-source tools was used, in order to ensure reproducibility. Representative tools and sample commands are:

- Hping3 (v0.9.0) — used inside the attacker virtual machines to generate volumetric and specially crafted packet floods sudo hping3 --flood -S -p 80 --rand-source 10.10.10.20.

- Nmap (v7.x) — reconnaissance from the attacker virtual machines (example): nmap -sS -A -p1-65535 10.10.10.0/24.

For each attack scenario, the attacker and victim virtual machine identifiers in EVE-NG and the precise start/stop timestamps were recorded in the SIEM system.

The purpose of carrying out these experiments is to evaluate the infrastructure's response in the event of an attack, to test various types of solutions to choose the best one and to test several types of countermeasures, which can improve the resilience of the network. To achieve this goal, several types of attacks were tested, both at the network level and through malicious files, various types of security solutions and different types of sources that can cause security incidents, both from the external and internal areas, accidental or intentional.

The experiments were designed to be repeatable by leveraging the EVE-NG snapshot functionality and a fixed running protocol. For each scenario, multiple independent runs were performed. Each run followed these steps:

- Bringing all virtual machines to a clean state.
- Starting SIEM services and agents and confirming agent connectivity for at least 60 seconds.
- Collecting normal traffic from monitored virtual machines.
- Initiating the attack sequence from the attacking virtual machines for a minimum duration of 60 seconds.
- Continuing monitoring for a period of 120 seconds post-attack.
- Archiving logs and exporting SIEM alerts for analysis.

After each run, the initial state of the virtual machines was restored to ensure isolation and eliminate side effects.

Practically, this section will include all the resources mentioned above, with the main purpose of having a unified exemplification of the concepts mentioned above and which are indispensable for a secure network.

The infrastructure presented in Fig. 1 was considered in order to be a simulated environment of a real network, with the purpose of testing various functionalities, in order to be able to observe the real implications. However, what should be mentioned is that the hardware and software resources owned by the real network are infinitely larger, compared to its simulated version. Nevertheless, this minor inconvenience is much less important than the purpose of the experiments carried out through this infrastructure.

Although during the experiments, several attack scenarios will be used, in a real case, analyzing the infrastructure in Fig. 1, the place of origin with the highest probability of achieving an intrusion is through a branch office. This unfavorable aspect is caused by the fact that, within them, there are more "relaxed" in terms of security measures and, sometimes, even the protections of certain resources are implemented in a more permissive and trust-based way.

This aspect is extremely unfavorable for the general security measures of the network because, in the event of an attack, the affected resources will increase exponentially. As a result, the attack can also extend to the Data Center or critical infrastructure area and can affect both the image of the organization and resources that should be operating in a constant manner.

Also, if we put aside the technical area, the huge impact of an attack will have financial implications, both for measures to prevent the spread, and for the remediation of the affected aspects (ransomware) or for the subsequent use of more effective protection measures.

The experiments were designed to be repeatable by leveraging the EVE-NG snapshot functionality and a fixed running protocol. For each scenario, multiple independent runs were performed. Each run followed these steps:

- Bringing all virtual machines to a clean state.
- Starting SIEM services and agents and confirming agent connectivity for at least 60 seconds.
- Collecting normal traffic from monitored virtual machines.
- Initiating the attack sequence from the attacking virtual machines for a minimum duration of 60 seconds.
- Continuing monitoring for a period of 120 seconds post-attack.
- Archiving logs and exporting SIEM alerts for analysis.

After each run, the initial state of the virtual machines was restored to ensure isolation and eliminate side effects.

Practically, this section will include all the resources mentioned above, with the main purpose of having a unified exemplification of the concepts mentioned above and which are indispensable for a secure network.

The infrastructure presented in Fig. 1 was considered in order to be a simulated environment of a real network, with the purpose of testing various functionalities, in order to be able to observe the real implications. However, what should be mentioned is that the hardware and software resources owned by the real network are infinitely larger, compared to its simulated version. Nevertheless, this minor inconvenience is much less important than the purpose of the experiments carried out through this infrastructure.

Although, during the experiments, several attack scenarios will be used, in a real case, analyzing the infrastructure in Fig. 1, the place of origin with the highest probability of achieving an intrusion is through a branch office. This unfavorable aspect is caused by the fact that, within them, there are more "relaxed" in terms of security measures and, sometimes, even the protections of certain resources are implemented in a more permissive and trust-based way.

This aspect is extremely unfavorable for the general security measures of the network because, in the event of an attack, the affected resources will increase exponentially. As a result, the attack can also extend to the Data Center or critical infrastructure

area and can affect both the image of the organization and resources that should be operating in a constant manner.

Also, if we put aside the technical area, the huge impact of an attack will have financial implications, both for measures to prevent the spread, and for the remediation of the affected aspects (ransomware) or for the subsequent use of more effective protection measures.

However, the most important implications are given by attacks in the NOC area or in the Data Center area because the people who have access to this data have much greater visibility over the entire infrastructure. Thus, attacks can be tested before the actual attack so as not to be easily detected or can be hidden within legitimate files, until they are detonated. Although it is less likely that users of these areas of the network infrastructure will initiate an attack (due to the level of training, strict policies or secure access to resources), if they do, the impact can be devastating.

The most common risks are data from compromised devices, intentionally or not, users unfamiliar with the cybersecurity area and the implications of security measures, phishing attacks that led to the attacker finding out key credentials, improperly secured Wi-Fi networks, configuration errors that can be exploited or simple credentials (admin/admin type).

The most likely attacks at the network level are infection with malicious files, data exfiltration, erroneous or apparently erroneous configurations that add additional, unauthorized routes, exploitation of weak security measures, use of firmware backdoors, DoS, unauthorized access to various types of resources, etc.

Some of these attacks will be presented in the following section in order to achieve a more realistic example of several types of attack scenarios and to be able to identify network vulnerabilities and the most effective methods of protecting them. The focus will be on network scans, identification of network flooding attacks or Denial of Service (DoS) attacks.

V. Results

In this section, we will consider several types of attacks, both network-based and file-based. The main goal is to visualize the attack response time and effectiveness within a fully virtualized system.

A. Network Scanning

Network scanning attacks are used to allow an attacker to view the essential characteristics of certain network devices or end-devices in order to create a network topology and discover various vulnerabilities. These vulnerabilities can be found both at the operating system level and at the exposed services level, and their exploitation is done in a unique way by each attacker.

In the experiments carried out at the network level, through the EVE-NG platform, the scans at its level were successful in detecting IPs and open ports in most of the devices tested (both for network devices: routers, switches, and for devices used by a normal user).

The main advantage was given by the SIEM system, which managed to capture most of these attacks, even though the network logs were extremely difficult to read and process due to the large volume of data. This aspect highlights the importance of monitoring and log capture measures both in complex and simple networks.

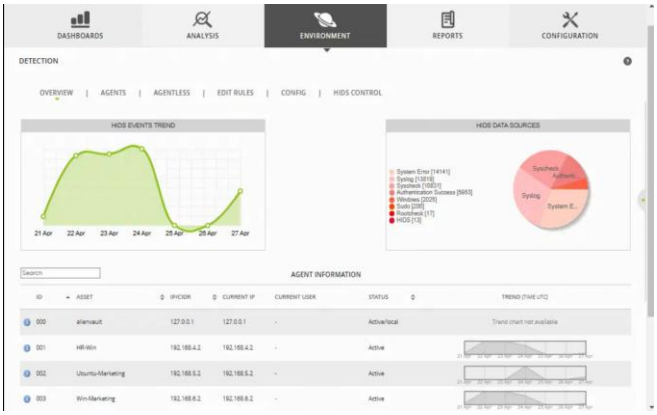


Fig. 2. Network architecture visualization mode at the SIEM system level.

Fig. 2 shows how the SIEM system analyzes the available agents and manages them efficiently, capturing their logs. As can be seen, the agents come in several categories, including various types of operating systems (Linux or Windows).

In the first test scenario, both attacks from the network area that includes the Marketing architecture and the network area that includes the HR part will be analyzed, both being contained in the same branch, as can be seen in Fig. 1.

Consequently, in this work scenario, both operating systems (Linux and Windows) were used as victim resources, and the attacks were generated using a Kali Linux virtual machine. This is not common in a normal network architecture, but the fact that the attacker machine is not a regular Ubuntu or CentOS does not impact the result at all because the attacks were generated using regular Linux commands and not with specific Kali Linux tools.

The first attack was carried out on the company's own network infrastructure, from the marketing area to the same network, to observe the available devices and open ports. This aspect led to the identification of certain attacks, which can be focused on those ports to destabilize the infrastructure. The way of carrying out the attack included running, through Kali Linux, the nmap 192.168.5.0/24 (local network) command.

Although the initial expectations were that the SIEM system would identify this attack and alert security administrators to this fact, it was concluded, based on experiments, that the Security Event and Information Management system did not even register the event, it was "lost" in the multitude of network logs.

The next work scenario targeted an attack from the Marketing area to the HR area, maintaining the same characteristics as in the previous case. Therefore, another, more aggressive attack was generated using the nmap -A 192.168.6.0/24 (distant network) command, which led to the identification of several features of the HR network area.

This time, the attack was recorded at the Security Information and Event Management system level, and the logs were displayed in a relatively short time at the dashboard level. Therefore, in this case, the SIEM system functioned effectively, recording the attack so that security administrators could

intervene to identify and isolate it. Following this type of event, a rule could be created to trigger an automatic action, which would isolate the traffic generated by an attacker workstation. Therefore, the resulting behavior consisted in the fact that the attacker station's IP is declared suspicious and the scan stops, the attack not being directed to the initial destination.

The way this type of attack is detected at the OSSIM level is through sensors that identify Nmap, Masscan or network scanning actions. These sensors usually generate an event in the dashboard that has a "Scan Detection" or "Potential Network Scan" alert. Following this event, alerts can be created to automatically block this type of traffic. In this case, the attack is not automatically blocked due to the fact that the test mode includes successful attacks, in order to test the efficiency of the Security Event and Information Management system in capturing logs from devices and displaying detected attacks or intrusions.

In the last scenario in the network scanning subchapter, the attack was carried out directly to the SIEM, in order to observe if and how it will prevent it. The attack was also generated through Kali Linux, as can be seen in Fig. 3 and was recorded in the SIEM system dashboard (Fig. 4). The important aspect is given by the fact that the attack led to temporary interruptions of the SIEM system, even if it consisted only of a trivial scan. Thus, even a small change, at the resource or configuration level, can lead to the total or partial interruption of log capture, leaving, at the same time, various types of attacks unnoticed and unverified. Practically, the latency in the response can indicate high CPU consumption for the analysis of each packet whose destination it is. In this subchapter, two different attack scenarios were presented in which the attacker was a Kali Linux device. The attacks were directed to a device within the network and to a device from another network, but in the same branch.

Although the initial attack was not recorded by the SIEM, the other two were recorded effectively and even alerts were displayed for security administrators.

```
root@kali:~# nmap -A 192.168.0.150
Starting Nmap 7.80 ( https://nmap.org ) at 2025-04-27 08:53 EDT
Nmap scan report for 192.168.0.150
Host is up (0.0031s latency).
Not shown: 996 filtered ports
PORT    STATE SERVICE  VERSION
22/tcp  open  ssh      OpenSSH 7.4p1 Debian 10+deb9u7 (protocol 2.0)
| ssh-hostkey:
|_  2048 3c:15:73:09:9f:2b:8e:ad:fb:f5:15:c6:99:01:4f:59 (RSA)
80/tcp  open  http     Apache httpd
|_http-server-header: Apache
|_http-title: Did not follow redirect to https://192.168.0.150/
443/tcp open  ssl/http Apache httpd
|_http-server-header: Apache
| http-title:  AlienVault OSSIM
|_Requested resource was session/login.php
| ssl-cert: Subject: commonName=alienvault
| Subject Alternative Name: email:system@alienvault.com
| Not valid before: 2025-04-19T18:20:11
|_Not valid after:  2035-04-17T18:20:11
|_ssl-date: TLS randomness does not represent time
| tls-alpn:
|_  http/1.1
514/tcp open  shell?
Warning: OSScan results may be unreliable because we could not find at least 1
 open and 1 closed port
Aggressive OS guesses: Linux 3.12 - 4.10 (94%), Crestron XPanel control system
 (90%), OpenWrt Kamikaze 7.09 (Linux 2.6.22) (90%), Linux 3.2 (89%), OpenWrt 0
.9 - 7.09 (Linux 2.4.30 - 2.4.34) (88%), OpenWrt White Russian 0.9 (Linux 2.4.
30) (88%), ASUS RT-N56U WAP (Linux 3.4) (87%), Linux 3.1 (87%), Linux 3.16 (87
%), HP P2000 G3 NAS device (87%)
No exact OS matches for host (test conditions non-ideal).
Network Distance: 2 hops
```

Fig. 3. Network attack including aggressive scanning of the SIEM system.

AlienVault HIDS: Windows error event.	2025-04-27 08:54:26	alienvault	N/A	Win-Marketing	Win-Marketing	2->2	LOW (0)	
AlienVault HIDS: Windows error event.	2025-04-27 08:54:26	alienvault	N/A	Win-Marketing	Win-Marketing	2->2	LOW (0)	
AlienVault HIDS: SSH insecure connection attempt (sca n).	2025-04-27 08:53:58	alienvault	N/A	Ubuntu-HR:56232	0.0.0.0	2->2	LOW (0)	
SSHd: Did not receive identification string	2025-04-27 08:53:57	alienvault	N/A	Ubuntu-HR	0.0.0.0:22	2->2	LOW (0)	
AlienVault HIDS: Windows error event.	2025-04-27 08:53:30	alienvault	N/A	Windows-HR	Windows-HR	2->2	LOW (0)	

Fig. 4. Displaying the attack in the SIEM system dashboard.

B. Denial of Service Attack Simulation

DoS (Denial of Service) or DDoS (Distributed Denial of Service) are some common network attacks that involve generating TCP connections with a large volume of data, which aim to temporarily block access to resources, affecting certain services. These types of attacks are usually focused on a specific device, and the impact consists of intentionally reducing its performance. In the favorable case where this system also has a backup connection, services will only be affected to a small extent, but in a real infrastructure, there are extremely few network areas that operate in this logic.

The essential countermeasures for this attack not to be successful consist of limiting multiple connections to a specific device, blocking various queries by firewall devices, equipment that operates in HA, devices that have integrated the concept of load-balancing, etc.

Moreover, if the attack comes from "inside", a good part of these measures becomes ineffective anyway because they have already been bypassed by the attacker. Therefore, it becomes imperative to be aware of the impact that an attack carried out from the internal area can have on the infrastructure.

To prevent these types of attacks, more access control measures must be implemented, SIEM systems that also integrate user behavior analysis (based on machine-learning) have to be considered and resource segmentation measures should be used.

The impact of this type of attack, from an experimental point of view, consisted in affecting network communications, temporarily interrupting certain essential services and degrading network performance. Moreover, the loss of certain packets at the level of a network of the size of the one in Fig. 1. is normal. This is true even if VPN connections are made between the Headquarters area and the branch area.

However, in the case of a DoS attack, the network's "defense" mode is to drop certain packets to cope with the overload. If the attack is long-lasting, at some point, the affected resource may even become unavailable, affecting even communication between certain parts of the network.

In the first working scenario, similar to the way the previous experiments were carried out, an attack will first be executed within the branch network, using the command:

hping3 -S –flood -V -p 80 192.168.5.2

The attack is recorded at the SIEM system level, but the alert is not clear enough to indicate the type of attack. However, multiple Windows errors indicate, for security administrators, abnormal behavior at the network level within the respective branch.

In the work scenario where the attack is centered on the Security Event and Information Management system (Fig. 5), the attack works in a similar way, but the SIEM system's reaction mode is extremely serious for the entire infrastructure because it becomes unavailable in the entire period of the attack.

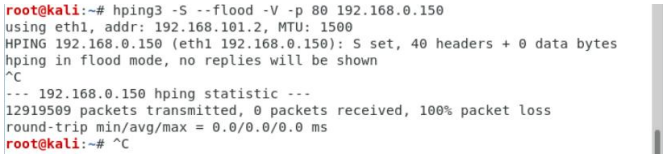


Fig. 5. DoS attack carried out on SIEM.

The immediate impact is given by the fact that, during this period, the Security Event and Information Management system does not record the attacks completely, and this aspect can completely affect the analysis network and the entire infrastructure of the organization. In practice, the SIEM system still receives logs from the devices in the network, but it can partially lose various events, and these are unrecoverable.

In this working scenario too, the EVE-NG test environment has proven important to simulate certain network behaviors in case of attack, allowing various types of important conclusions to be made, through the experiments carried out.

C. Malicious File Attacks

During the experiments, network intrusions through malicious files were also considered. The way in which these files came to penetrate the network was not included in the experimental data because it is not the scope of this work.

Consequently, also through the EVE-NG virtualized environment, a realistic scenario was created in which attacks originating from the internal area of the network were simulated and which are masked in legitimate files, which are executed due to lack of knowledge of their real nature. This type of attack is used to compromise the security of systems and to steal certain data about users or about the organization. The goal is to obtain various financial facilities or to steal various databases or ideas, for the purpose of reselling them or for the purpose of blackmail.

The test network included all the elements in Fig. 1, and the attack scenarios were triggered manually or automatically, through malicious files, transferred via FTP or downloaded from the public resource area, by a legitimate user. The attacks in this test scenario, however, came from several different parts of the network, in order to test both the intervention from the NOC area and the intervention from other branches, compared to the one tested in the previous scenario.

Basically, in this test scenario, the entire network was intended to be involved, but due to the multitude of logs, the SIEM system worked inefficiently, having moments of crashing or difficult event recording. Thus, for efficient detection, in certain experiments (based on more critical files such as ransomware), unused resources were closed. This way of working implies a more efficient use of resources in the test scenario through EVE-NG, but in a real test scenario, this is almost impossible, the purpose of the SIEM system being, mainly, to capture logs from all devices used in the network. Therefore, in this test scenario, the proportion of effective tests in EVE-NG, whose conclusions can be directly translated into the real working environment, was not respected. However, the way the SIEM system works, the way alerts appear and are customized at their level, and the way scripts that use these alerts are identified, can be implemented based on these tests.

The biggest disadvantage of these experiments in EVE-NG and the real environment is that, although the network in Fig. 1 is a large-scale network, it is nowhere near the size of a real (production) network. However, the resources used by the SIEM system to process all these logs are quite high, but the processing is not that efficient. In addition, considering the fact that the volume of data processed is insignificant compared to the real volume of data, the performance analysis on the test environment cannot fully reflect the multiplicity of the real environment.

Considering that in EVE-NG, the resources used do not have access to actual internet resources, in some experiments, Kali Linux devices were also included. The explanation for their use, compared to the use of regular Ubuntu devices, is that the former have more pre-installed libraries. An example, in this case, is the SSH (Secure Shell Protocol) service, which, in Kali Linux, has both the server and client sides installed, while in Ubuntu, it only has the client side. Therefore, Kali machines are considered regular users and agents of the network and not attacker machines.

The attacks were analyzed dynamically, including monitoring of CPU and memory resources and detection by antivirus on the victim host. Also, in some experiments, Wireshark tool was included.

One of the most relevant tests was the one regarding a ransomware attack, implemented by executing an .exe file, from the public resources area. Therefore, within this scenario, two test modes were implemented: a ransomware attack originating from the NOC area and an attack originating from one of the branches. The difference between the two types of attack was given by the affected operating system, focusing on the comparative analysis of a ransomware attack in a Linux-based operating system and in a Windows-based operating system.

The main purpose of the two types of attack was to visualize how the SIEM system registers the intrusion and to observe the impact that this malware attack has on network resources (both on the affected workstation and on its adjacent resources: routers, switches or other workstations).

The aspects observed after execution recorded typical behaviors of a ransomware file, this file scanning the disk within Windows and Linux systems and encrypting all accessible files,

in an extremely short time. Each file also received a different extension and a .txt file was created, highlighting an email and an amount through which the affected data can be recovered. It is mentioned that the behavior of the ransomware file was not detected by the antivirus system at the time of its download, remaining undetected until the time of execution. The effect on the infrastructure was irreversible, the files being inaccessible, without a backup solution.

The SIEM system did NOT manage to detect the attack in time (at the time of downloading the malicious file on the client's workstation), thus making the entire network vulnerable. And, at the time of the attack, the intervention on the resources was already far too late to be implemented effectively.

In the first experiment, a ransomware attack was carried out on the Ubuntu operating system, in branch 3, which contains the DEV department and the IT development area.

During the ransomware experiment, the execution of the .elf file [18] was attempted. However, this file required additional dependencies and dynamic libraries. Therefore, for the file to function properly, these dependencies had to be installed. Not having direct access to internet resources and having various problems with the offline installation of the necessary dependencies, this attack did not function properly, in a final version.

Within the SIEM system, the attack generated in the Ubuntu workstation is not visible, the logs not recording the execution of this file, neither when it did not have the necessary dependencies nor after they were installed in an offline mode, and the attack was run again (Fig. 6).

Fig. 6. Ransomware attack file execution from branch 3.

This aspect is quite complicated, given the nature of the malicious file executed because, even if, in the end, its execution did not affect the operating system, the fact that the file is not even registered at the Security Event and Information Management system level requires the existence of additional protection measures to compensate for this faulty way of working (by implementing IDS/IPS solutions or through application-type firewalls, which record and restrict such behaviors). The next working scenario included the same way of working, but within the Windows operating system. This time, the file chosen was .exe [19] and it worked properly, and the attack was successful, as can be seen from Fig. 7-9, which presents the entire evolution of the attack. In this case, contrary to the previous example, the Security Event and Information Management system detects malicious files and displays alerts at the graphical interface level.

Although, even in this case, the alerts are not very conclusive for the security team to know that there is a ransomware attack on the network, the multitude of errors that appeared in the dashboard require additional verification. An important aspect to note is that, although there was routing at the network resource level and the affected station was connected to the external environment (including the SIEM system, being its agent), the ransomware attack did not affect other resources on the network.

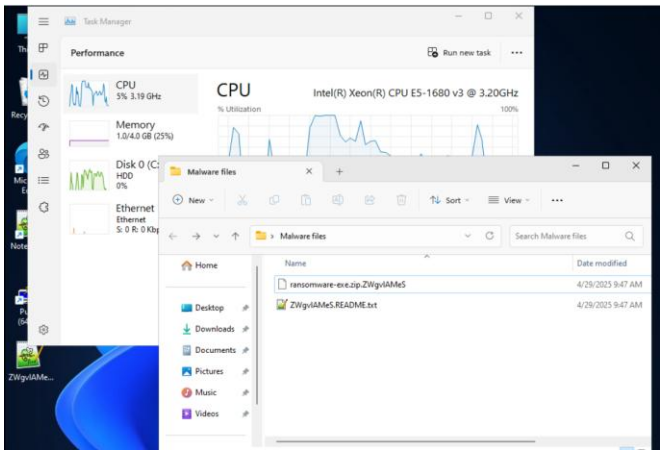


Fig. 7. Ransomware attack on windows operating system.

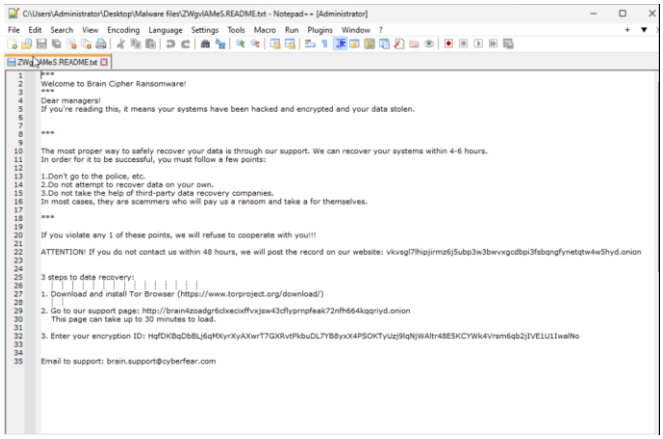


Fig. 8. Detailed README.txt of ransomware attack.

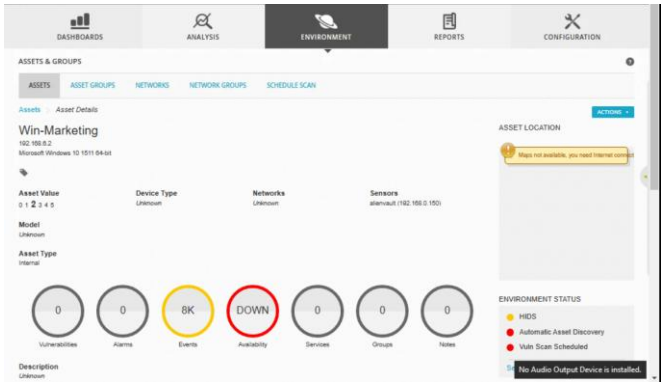


Fig. 9. The victim client station becomes inaccessible to SIEM after the ransomware attack.

Another extremely important aspect is the fact that, following the attack, the affected client station also became unavailable at the Security Information and Event Management system resource level, appearing DOWN, although it continued to function, only that its files had been encrypted. Therefore, it is considered that the ransomware affected the services, also

blocking communication with the manager, but did not specifically target the security agent.

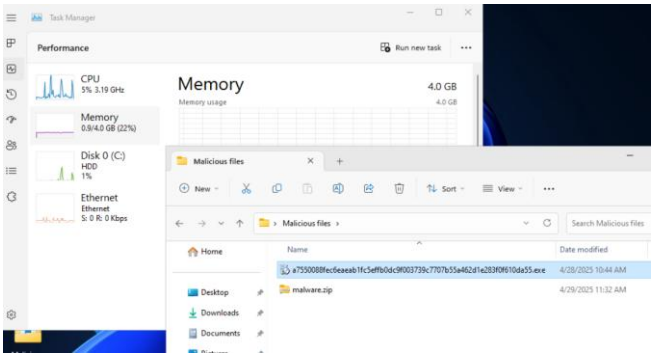


Fig. 10. Malware attack file execution from branch 3.

In a similar scenario, an attack with a malware file [20] on a Windows node was also considered, simulating another compromise, from NOC area (Fig. 10). The impact was given by the fact that activity was detected at the antivirus level, and the system resources (CPU, memory) were affected (CPU – 100%). At the level of the system that ensures Security Information and Event Management, the attack was detected, but the alerts are different, compared to those displayed previously, but still not clearly expressed enough to easily identify that it is a malware attack (Fig. 11).

In the present case, propagation in the virtual network is possible, due to the malicious nature of the file, but calls to the internet are not possible because the victim virtual machine does not have outbound access from the network. Thus, the file can see the LAN area, but, without a source of command and control, in the test environment used and during the experiments, it did not try aspects of lateral movement.

Fig. 11. Displaying the malware attack in the SIEM system dashboard.

The malware acts by running and modifying registries (from subsequent tests, it remains on that virtual machine even after reboot), but is limited to the actions performed, being valid only within the virtual workstation. To verify that the malware was propagated in the LAN area, some Wireshark and Process Monitor captures were performed on the victim station, observing both the malware scan logging and the behavior on processes.

In order to simulate a real internet environment, similar to some experiments that are carried out in real sandbox environments (not those based on EVE-NG), it would have been necessary to use tools such as FakeNet [21].

It can simulate network services and responses so that the malicious software does not notice the fact that it is a test, virtualized environment and reveals its real behavior. In this way, domains or command-and-control servers can be identified. In EVE-NG this simulation was not possible because, although a separate workstation was used, acting as a gateway, the malicious file does not adapt its behavior.

In this study, our goal was not to calculate statistical detection indicators such as false positive or false negative rates, but to verify whether the SIEM system (OSSIM) successfully recorded the occurrence of specific attacks in the environment emulated by EVE-NG. For each experiment, attacks were generated using hping3 (DoS and ICMP/TCP floods), Nmap (port reconnaissance and scanning), and two real malware samples executed in virtual machines. The outcome for each scenario was evaluated based on the detection and recording of the event by OSSIM (Recorded = Yes/No).

Table II provides a summary of all scenarios, indicating which attacks produced SIEM alerts.

TABLE II. SIEM EXPERIMENT RESULTS

Scenario	Tool/Sample	No of Runs	Recorded in SIEM	Alert Type
Port Scan	Nmap	2	Yes	SSH insecure connection attempt
DoS	hping3	2	Yes	Traffic Anomaly
Malware Sample 1	Ransomware1	1	No	-
Malware Sample 2	Ransomware2	1	Yes	Registry integrity checksum changed
Malware Sample 3	Malware	1	Yes	Registry integrity checksum changed

Another aspect that should be considered in this section is false positives. Given that in a network environment like EVE-NG there are a lot of devices generating traffic in parallel, a detailed explanation of all these values, which can appear at the SIEM level, is necessary.

The first one is about legitimate high-volume traffic because, sometimes, this traffic can generate attack-like patterns, but some VMs are doing normal operations (updates, backups, scans).

Another one can be when several VMs communicate in parallel and the SIEM may interpret aggregated traffic as suspicious.

The third one can be generated by multiple concurrent updates or temporary files can be flagged as suspicious.

Some alerts may correspond to benign parallel activity in the EVE-NG environment, representing potential false positives inherent in multi-device traffic aggregation. These were identified and differentiated from alerts generated by attacks by controlling the timing of the attacking virtual machine and even isolating a part of the network, when the attack was generated, in order to observe its impact on the SIEM system.

VI. Discussion

The experiments used in this chapter included several types of attacks generated at the level of a test infrastructure. Therefore, multiple attack scenarios were carried out, which included both network scanning attacks, DoS attacks and malware infections.

Considering that it is a test environment, EVE-NG cannot cover a real infrastructure, most of the resources being limited. Consequently, there are several cases in which the infrastructure operates slowly and, considering that a SIEM system also operates at its level, even storage will be even slower than in a real environment. The OSSIM Security Event and Information Management System requires substantial resources to function optimally. Being also an infrastructure in the EVE-NG environment, it can sometimes be overwhelmed.

Attacks carried out on this network are quite suffocating for SIEM, even if the hardware resources are sufficient because it must process an extremely large number of network logs, from all the devices. Moreover, during the attack experiments, a good part of the remaining devices was shut down, so that the logs could be tracked as easily as possible and so that the analysis would not overload the available resources. Event processing at the SIEM system level includes each network log being inserted into a database, being correlated with other events in the network and being displayed in a graphical interface. Thus, the processing time is extremely long and, sometimes, alerts appear with delays.

The SIEM system is built for the analysis of normal traffic events. The intrusion detection part is an additional one and, due to the fact that an EVE-NG system is also used, in which it is integrated, this area is sometimes inefficient. Although the Security Event and Information Management system is optimized to function as a network device in EVE-NG, it is not used to the massive volume of traffic, as in the case of previous attacks (directed to it). Therefore, when it detects an attack on it, it has two types of behavior:

- It can ignore the attack and continue recording events from legitimate devices.
- It may freeze, delay event display, or become unavailable.

In the work scenarios, it was required that the attacker know the IP of the SIEM system and that it not be protected by any firewall or other resources, but in a real work environment, this aspect is quite impossible for a branch user. It is not, unfortunately, impossible for a user working in the NOC area who could have access to such resources. However, most of the time, routing resources to the datacenter infrastructure level (where the Security Event and Information Management system is located) is not so accessible. As prevention methods, these would include separating the infrastructures at the datacenter level through hardware and software firewalls and segmenting the network in the most efficient way possible. Also, another urgently needed method is the implementation of domain policies that include limiting user access to tools such as nmap or to sensitive network resources.

A particularly important aspect is the ease with which this type of file can penetrate the network, in the absence of policies for their automatic detection. Therefore, without this type of monitoring and blocking policies, the files end up being executed, without the user being able to imagine their real intention. Therefore, inspection within the SIEM, but also the presence of an antivirus becomes essential elements for securing the environment in the endpoint area. Following the execution of the files, they had various behaviors, but most of them fell within the aspects mentioned in the chapter on static and dynamic analysis of malicious files. However, if we also take into account the impact aspects at the network level, it is found that they tried to generate various illegitimate connections, created process delays and resource consumption and various types of temporary blockages.

The disadvantage is that, in some cases, malicious files passed undetected from one computer to another, even if their activity was monitored by the SIEM system. Basically, the logs related to this type of transfer went unnoticed, the alert being created only at the time of execution. The fact that the SIEM system could not automatically block this type of file required manual analysis, thus implying the possibility of extending the attack to the entire network.

Simulating these attacks in the EVE-NG virtualization environment allowed highlighting the easy way with which malicious files can completely compromise a network, observing the behavior of the network in the event of an infection and testing the real infrastructure, without implications in the production area. However, a troublesome feature in the experimental area consisted of the direct implications that certain agents had when disconnecting from the manager's area. Therefore, considering that most of the experiments were carried out with closed network parts, when reconnecting them, there were various desynchronizations between the agent and the manager. Basically, the time in the agent device was no longer correlated with the time in the manager area, and the logs, including at the dashboard level, were not recorded. This way of working has been encountered previously in the real working environment, and, in both cases, it produces serious issues because the attacks are not recorded properly, and the SIEM system does not alert administrators in any way about this aspect. It is obvious that, without integrated security solutions and correct network segmentation, even in a controlled environment, attacks can have devastating effects. The experiments carried out have demonstrated that a simple file executed by a user with limited rights can serve as an entry point

for ransomware, backdoor or other forms of malware that escalate privileges and compromise the entire infrastructure.

In most cases, the SIEM system proved to be ineffective. Although the alerts provided raised the attention of network or security administrators regarding abnormal behaviors, they were quite vague, with no clear demarcation between the types of files that affected various operating systems and the way the alerts were displayed. Moreover, in some cases, the alerts displayed at the Security Information and Event Management system level are identical, although the types of intrusion files are different. If it is also considered the fact that, in some cases, the SIEM system did not detect a certain malicious behavior at all, we find the permanent need for additional intrusion detection measures (separate IDS/IPS systems).

The proposed method has potential applications that extend beyond the quality assessment of virtual devices, in a fully virtualized environment, such as EVE-NG. Therefore, the proposed method is considered useful for making informed decisions, in an environment that simulates a production environment, based on cybersecurity strategies that can adapt to multiple types of attacks and for multiple types of network devices or operating systems. This aspect also introduces efficient methods for learning and adapting the security policies of organizations, through various work scenarios, which can help to understand the cyber threat landscape.

Therefore, the main contributions of this work are the fact that a proprietary network architecture is proposed, divided into several different blocks, similar to a real network environment, which also involves resource segmentation. Also, the implementation of a SIEM system is considered to automatically detect network threats, operating in an agent-manager system, and including a centralized dashboard (where various alerts can be graphically viewed), like a real environment. Through this system, various logs can be collected from network equipment and workstations and automatic responses to intrusions can be created.

In addition, various work scenarios are tested, including both malicious files and network-level attacks, in an environment that is not specific to these types of actions. Also, the efficiency of these attacks is evaluated, both at the level of the network's own resources and at the level of the other parts of the network (for detecting lateral movements and for testing how an attack works at the level of the entire network).

Moreover, the EVE-NG working platform is evaluated, under the conditions of a complex network architecture and malware detection, at the SIEM level, although the platform was not created for these purposes, but only for common network tests and networking knowledge training.

To summarize, the main contributions of this paper are as follows:

- We designed and implemented a controlled cybersecurity experimental framework using EVE-NG, demonstrating its potential to model both network attacks and, more superficially, real malware attacks, in a secure environment.
- The feasibility and limits of using a non-dedicated network simulator for security research were assessed, providing insights into its adaptability and practical constraints.
- Experimental results were presented illustrating the behavior and network impact of common attack types.
- The potential role of EVE-NG as an accessible and cost-effective platform for cybersecurity training, testing, and education was highlighted.
- The response of various emulated systems to attacks or attack files was tested, in order to be able to formulate various preliminary conclusions to define their influence in production environments.

These contributions aim to reduce the gap between theoretical cybersecurity studies and practical experimentation, supporting both academic and practical advances in the field of network security.

Although the proposed work has multiple advantages, the imminent limitations of any system must be taken into account. The first one is that malware samples can identify if the sandbox environment is virtualized and can modify their behavior depending on this characteristic. Therefore, it would be much more appropriate for this simulation to be performed in a real test environment, not a virtualized one. However, most of the time, this is not possible because the malware execution can lead to partial or total degradation of the test environment.

The second one can be considered by the limited number of experiments because the specific behaviors of a certain type of malware or attack cannot be extracted except after successive runs of several experiments with similar malware samples or from the same family. Such work scenarios are considered in the following tests.

Although the experiments are performed in both Linux and Windows operating systems, the third limitation is that the mobile area is not covered in this work. Other operating systems are also considered in the following experiments.

Another limitation, the fourth, is that the alerts displayed at the SIEM dashboard level are not very explicit and can be confusing if similar malicious samples (trojan and malware) are used. This can make it difficult to quickly interpret security events and prioritize incidents. In further developments, increased attention will be paid to customizing alerts, so as to provide more detailed and useful information. It will also be considered to evaluate other types of SIEM systems, such as Wazuh [22], to test their ability to generate more intuitive and informative alerts.

The last one can be considered the fact that the test scenarios included experiments were performed in short period of time (up to one hour) in order not to degrade the test environment. This aspect is considered a limitation because some malicious files or some attacks become more effective after a certain period in which they "adapt" to the victim environment.

This type of test system is effective for supporting conclusions similar to the real environment because, in a virtualized environment, there are constraints on various types

of resources. Thus, the resulting ideas can only be as a possibility of similar action of a malware in a real environment, most of the time, in the second case, the effect being much more intrusive. The most important measures for protecting the analyzed network are:

- The existence of a SIEM system, to be able to capture various network logs and to be able to effectively detect malicious executions or attacks from the internal area of the network.
- Implementation of much stricter access control policies, in order to limit users from easily creating attacks.
- Segmentation of various parts of the network, to prevent the penetration of illegitimate traffic to critical resources.
- Constant updates of antivirus systems, in order to have access to the latest versions of malicious file signatures.
- Constant updates of antivirus policies and systems at the endpoint level, to block the execution of malicious files by users.
- Implementation of secure password policies, which are constantly updated.
- Performing constant backups, in environments different from the usual ones, to reduce the effects of ransomware attacks.
- Performing experiments in virtualized environments, to track the weak characteristics of the network.
- Educating users through concrete examples, to increase awareness of the dangers.
- Verifying, through static and dynamic analysis methods, the malicious files that have penetrated the network, to fully understand their behavior and the network vulnerabilities (if they have penetrated the systems by downloading or through FTP servers and have not been detected by antiviruses or SIEM, a problem is indicated at the level of the policies implemented within these solutions).
- Using, besides SIEM system, additional IDS/IPS solutions.

This study focused on evaluating the ability of OSSIM to record and alert on network and host-based attacks in an EVE-NG environment. Quantitative evaluation of mitigation measures, such as firewall rules, traffic filtering, or SIEM correlation tuning, was not performed and is beyond the scope of this work. But, the most important ones, for all network environments, can be:

- Firewall rules or access control lists (ACLs) – that can block Nmap scans, hping3 floods, SSH brute-force attempts.
- Rate limiting or traffic shaping – that can mitigate TCP/ICMP floods (hping3).
- Segmentation – this kind of mitigation was considered even in the network topology and it's advantage can be the fact that it limits spread of malware and reduces alert noise in SIEM.
- Endpoint firewalls – that can prevent malware scanning other hosts.
- Whitelist trusted hosts – that can reduce the impact of a new malicious host.
- Logging and backup – that can retain logs to analyze attacks, false positives, and potential missed events.

In this work, it was proven that EVE-NG is an efficient test environment, but it has its limitations when the network architecture is more complex or when malicious aspects are introduced, which must be processed in parallel with the usual activities of the network devices.

But, citing again the work [13], no test environment (hardware or software) can contain, include or be impenetrable to any type of malicious software. The goal is to learn in as much detail as possible the characteristics of these files, in order to be able to offer, at any time and under any conditions, the most efficient defense and a significant reduction of their impact.

VII. Conclusion

In this work, the focus was on simulating a computer network similar to a real environment, which included both network devices comparable with those currently in use, as well as workstations with Linux or Windows operating systems. This network was created highlighting the idea of resource segmentation and included a central office and remote branch environment. Thus, five major components of the network were imposed: the datacenter area, which included server resources of various types and the network's SIEM system, the NOC area, which ensures continuous monitoring of resources, and the remote branch area, which are related to the resources used in the datacenter area and which are constantly monitored through the NOC area.

The routing method, as well as the services running at the network level, were not central to this work, the emphasis being placed on the analysis of intrusions and network traffic, in the usual way.

In the experiments carried out, two different types of network traffic analysis were considered: network traffic analysis in the presence of an attack at its level and network traffic analysis in the case of an attack by a malicious file. The malicious files used were downloaded from MalwareBazaar resource and are available online.

The conclusion of these experiments is that, although EVE-NG is a good test environment for network administrators, when various types of experiments similar to the real environment are desired, if attacks are also considered, it becomes quite inefficient. Although it also implements methods based on SIEM systems, which are usually capable of detecting these attacks, experiments show that it can only partially simulate real-world analysis.

However, from the experimental area, one can conclude both benefits in the area of learning with this test environment, as well as the identification of key parameters in the analysis of data

networks. Thus, it is found that network traffic analysis must include several methods and solutions (static analysis, dynamic analysis, monitoring and constant logging through SIEM and analysis with machine learning methods) ensuring redundancy when one method fails. In this sense, if in some cases there is a possibility of duplicating the solutions (main SIEM and backup SIEM) or of performing a more complete analysis, using solutions that, apparently, offer the same results, the composition of a network environment as secure and robust as possible is all the better defined.

This study highlights that EVE-NG provides a secure and repeatable environment for examining malware behavior, offering a practical bridge between theoretical research and hands-on cybersecurity training. By enabling controlled experiments in a virtualized network, it allows both academic researchers and industry professionals to analyze attack patterns safely, supporting reproducibility and experiential learning in cybersecurity practices.

In future work, other types of malicious files will be considered, and the analysis will be performed on other types of operating systems that can be modeled in EVE-NG. Also, research will focus on incorporating Wazuh, to enable real-time behavioral analytics. These enhancements aim to improve threat identification capabilities, allowing the system to dynamically recognize unusual activity and provide timely, actionable insights for cybersecurity monitoring and response. The ultimate goal is to create a robust database containing more recent malware samples, which can be adapted to current security systems and learn as much as possible about malware characteristics, regardless of the test environment used.

REFERENCES

[1] S. U. Shaukat, S. Khan, and S. Parkinson, "A Review on Multi-Step Attack Detection," *IEEE Access*, vol. 13, pp. 161779-161805, Sep., 2025, doi: 10.1109/ACCESS.2025.3607497.

[2] D. Omand, "Social Media Intelligence (SOCMINT)," The Palgrave Handbook of Security, *Risk and Intelligence*, pp. 355–371, Jul., 2017, doi: 10.1057/978-1-137-53675-4_20.

[3] EVE-NG, "The Emulated Virtual Environment for Network, Security and DevOps Professionals, " Accessed Aug., 2025, available: https://www.eve-ng.net/

[4] A. Marefat, A. A. M. Nishar, and A. Ashok, "Text2Net: Transforming Plain-text to a Dynamic Interactive Network Simulation Environment," SoutheastCon 2025, Concord, NC, USA, pp. 625-630, 2025, doi: 10.1109/SoutheastCon56624.2025.10971486.

[5] G. Önal and M. Güven, "Enhancing Dynamic Malware Behavior Analysis Through Novel Windows Events With Machine Learning," *IEEE Access*, vol. 13, pp. 153937-153958, 2025, doi: 10.1109/ACCESS.2025.3604979.

[6] A. K. Roy and A. Kumar Khan, "Performance Degradation in Wireless Mesh Networks via External and Internal Attacks," 2019 2nd International Conference on Innovations in Electronics, Signal Processing and Communication (IESC), Shillong, India, pp. 258-262, 2019, doi: 10.1109/IESPC.2019.8902374.

[7] Malware Bazaar, "A Repository for Malware Samples," Accessed: Dec. 2, 2024, available: https://bazaar.abuse.ch

[8] Q. Liu, K. Bao, W. U. Hassan, and V. Hagenmeyer, "HADES: Detecting and Investigating Active Directory Attacks via Whole Network Provenance Analytics," IEEE Transactions on Dependable and Secure Computing, 2025, doi: 10.1109/TDSC.2025.3611866.

[9] A. Jony, M. N. Islam and I. H. Sarker, "Unveiling DNS Spoofing Vulnerabilities: An Ethical Examination Within Local Area Networks," 2023 26th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, pp. 1-6, 2023, doi: 10.1109/ICCIT60459.2023.10441649.

[10] A. Jony, M. N. Islam and R. A. Talukder, "A Secure Token-Based Approach for DHCP Client Authentication and Replay Attack Prevention," 2024 27th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, pp. 855-860, 2024, doi: 10.1109/ICCIT64611.2024.11022024.

[11] W. Qin, "Design and Implementation of a Cyber Rang with Emulated Network Security Devices," 2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT), Jilin, China, pp. 1526-1531, 2023, doi: 10.1109/ICCECT57938.2023.10141361.

[12] Maroš HARAHUS, Matúš ˇCAVOJSKÝ, Gabriel BUGÁR, and Matúš PLEVA, "Interactive Network Learning: An Assessment of EVE-NG Platform in Educational Settings", *Acta Electrotechnica et Informatica*, Vol. 23, No. 3, pp. 3–9, 2023, doi: 10.2478/aei-2023-0011

[13] Vasani Vatsal, Amit Kumar Bairwa, Sandeep Joshi, Anton Pljonkin, Manjit Kaur, and Mohammed Amoon, "Comprehensive Analysis of Advanced Techniques and Vital Tools for Detecting Malware Intrusion," *Electronics* 12 (20), 4299, 2023, https://doi.org/10.3390/electronics12204299

[14] Wireshark Foundation, "Wireshark: Network Protocol Analyzer," accessed: Mar. 6, 2025, available: https://www.wireshark.org/

[15] VirusTotal, "VirusTotal," accessed: Mar. 18, 2025, available: https://www.virustotal.com/gui/home/upload

[16] Google, "Google Rapid Response (GRR) Incident Response Framework," accessed: Mar. 21, 2025, available: https://github.com/google/grr

[17] AT&T Cybersecurity, "OSSIM: Open Source Security Information Management," accessed: Mar. 22, 2025, available: https://cybersecurity.att.com/products/ossim

[18] MalwareBazaar Database, "carolina-happy-mike-friend," accessed: Mar. 23, 2025, available: https://bazaar.abuse.ch/sample/f91d55e43648111bbfc06c0dba5ba80eb90b0ebc2a6dab6691f25aacc5c0ee2e/

[19] MalwareBazaar Database, "uranus-hamper-kilo-bacon," accessed: Mar. 23, 2025, available: https://bazaar.abuse.ch/sample/2d04d802438ae93b095acfbb87cf5760bfaf1bbd300a609d6941a6861bcc68a7/

[20] MalwareBazaar Database, "fanta-carbon-fruit-cold", accessed: Mar. 23, 2025, available: https://bazaar.abuse.ch/sample/a7550088fec6eaeab1fc5effb0dc9f003739c7707b55a462d1e283f0f610da55/

[21] Mandiant FLARE (FireEye), "FakeNet-NG — Next-Generation FakeNet (tool)," accessed: Mar. 23, 2025, available: https://github.com/mandiant/flare-fakenet-ng

[22] Wazuh, "Wazuh — Open Source Security Monitoring," accessed: Mar. 24, 2025, available: https://wazuh.com/

[23] Splunk, "SIEM: Security Information & Event Management Explained" accessed: Oct. 16, 2025, available: https://www.splunk.com/en_us/blog/learn/siem-security-information-event-management.html

[24] IBM, "IBM QRadar", accessed: Oct. 16, 2025, available: https://www.ibm.com/products/qradar

Predicting Stock Market Performance Based on Sentiment Analysis of Online Comments

A Case Study of China's Highest Market Cap Stocks

Wenhao Suo[1]*, Tongjai Yampaka[2]*
Chakrabongse Bhuvanarth International Institute for Interdisciplinary Studies (CBIS),
Rajamangala University of Technology Tawan-OK, Bangkok, Thailand[1, 2]
School of Management, Guangzhou Huashang College, Guangzhou, China[1]

***Abstract*—In China's retail-focused stock market, the influence of social media sentiment during off-hours on the next day's opening price has received limited attention. This paper takes Kweichow Moutai—a leading Chinese company with substantial market capitalization—as the research sample. It gathers investor commentary data from financial platforms, and uses natural language processing tools (SnowNLP) to develop a multidimensional sentiment index (including average sentiment score, positive ratio, and sentiment volatility). By integrating this index with stock trading data and macroeconomic indicators, this study designs a dual-channel LSTM model: one channel for market technical features (e.g., price, volume) and the other for sentiment features, aiming to analyze the impact of off-hours sentiment on opening prices. Empirical results indicate that overnight sentiment has significant predictive power for the next day's opening price; meanwhile, sentiment transmission is asymmetric, making predictions more challenging in declining markets. Additionally, high-frequency sentiment data significantly outperforms low-frequency data in market prediction accuracy. This research expands the understanding of how investor sentiment influences the market over time, providing practical insights for market participants to develop effective strategies and manage risks.**

***Keywords*—Investor sentiment; non-trading hour sentiment; social media comments; dual-channel LSTM**

I. Introduction

In recent years, China's stock market, the world's second-largest, has surpassed one trillion yuan in combined market capitalization for the Shanghai and Shenzhen stock exchanges. Structurally, the Main Board, ChiNext, and the STAR Market form a multi-tiered capital market system, providing businesses with a range of financing options while also better serving investors' increasingly diverse asset allocation needs. However, compared to mature markets, the Chinese stock market still has unique structural features. A trading environment primarily driven by retail investors yields high volatility, high turnover, and a sentiment-driven market. Individual investors significantly influence both the number of accounts and trading volume, thereby affecting market pricing efficiency and volatility. Given these features, investor sentiment—a hallmark of irrational behavior—is increasingly seen as a significant factor driving market volatility.

In recent years, the widespread use of social media has created a new data source for researching investor sentiment. Platforms such as Weibo, stock forums, and Xueqiu have become key channels for investors to gather information, share opinions, and interact, producing a vast amount of user-generated content (UGC) closely linked to financial markets. This unstructured text not only captures investors' responses to company developments, policy changes, and macroeconomic trends in real-time but also offers high frequency and broad coverage, making it a valuable resource for sentiment measurement studies. While much research in traditional finance has examined how sentiment influences the market from a behavioral finance perspective, Aggarwal (2022) notes that, despite over fifty years of investigating market sentiment, a lack of standardized definitions and measurement tools remains, leading to significant variation in empirical findings. Therefore, it is essential to revisit the use of sentiment variables in finance within a psychological framework to enhance modeling accuracy and predictive capabilities.

In the context of social media, Bollen, Mao, and Zeng (2011) [1] first attempted to construct sentiment time series using Twitter content. They discovered that sentiment dimensions such as "calmness" significantly enhanced the forecast of the Dow Jones Industrial Average (DJIA), confirming the forward-looking influence of macro-level collective sentiment on financial markets. Yang and Mo (2016) [2] also showed that extreme news or social media sentiment can act as leading indicators of market returns and volatility, with even greater predictive power when sentiment varies significantly.

The real-time and scalable nature of social media data makes it a valuable tool for financial research. Corea (2016) [3] found in their study of the technology industry that behavioral variables such as tweet volume and interaction frequency are more predictive of stock prices than tweet sentiment polarity. Li et al. (2017) [4] proposed the SMeDA-SA method to extract sentiment signals from massive tweets, achieving highly accurate predictions of individual stock fluctuations and further validating the feasibility of social platforms as proxies for investor sentiment. Liu et al. (2015) [5] also explored the role of social media indicators in predicting stock co-movement, finding that companies with active Twitter accounts tend to exhibit more consistent stock price movements, revealing the power of social media to shape industry structural sentiment.

*Corresponding authors.

Advances in sentiment recognition technology have also expanded its applications in the financial sector. Man et al. (2019) [6] systematically reviewed the development of financial sentiment analysis (FSA), from early dictionary methods and traditional machine learning to deep learning models such as LSTM and Transformer, providing a diverse range of tools for financial sentiment modeling. Wang et al. (2022) [7] developed a market-neutral strategy based on LSTM and experimentally confirmed that the model exhibits high accuracy and stability in extracting sentiment from individual stocks. Koukaras, Nousi, and Tjortjis (2022) [8] utilized the XGBoost method, while Kumar et al. (2022) [9] employed a fuzzy neural network (FNN) approach. Both methods combined tweet sentiment with historical market trends to achieve effective predictions of the Dow Jones Industrial Average and the stock prices of several companies. In addition, Hasselgren et al. (2023) [10] proposed visualizing the sentiment scoring system, providing a feasible tool prototype for investment decisions.

However, while existing research has produced valuable insights in social sentiment modeling and market forecasting, it has mainly concentrated on the real-time relationship between sentiment and market variables during trading hours. It has paid little attention to the mechanisms that influence sentiment beyond trading hours. Traditional financial theory posits that prices are primarily determined during trading hours through the order book, with information outside of trading hours being indirectly reflected through overnight orders. Nonetheless, social media platforms often generate a large volume of emotional expression between the market close and the next day's opening, especially during emergencies, policy announcements, or the release of financial reports. These sentiments can potentially shift investor expectations in advance, affecting the opening price on the following day.

In response to this research gap, some scholars have begun to explore this area: Ranco et al. (2015) [11] found, based on the event study method, that there is a significant correlation between Twitter sentiment peaks during non-trading hours and the next day's cumulative abnormal returns; Fraiberger et al. (2021) [12] pointed out that international news released at night significantly affects cross-border capital flows and global stock price returns, highlighting the importance of non-trading period information in the international market [12]; Goutte et al. (2023) [13] also confirmed, in the context of ESG investment, that news sentiment indicators during non-trading hours show strong robustness in market return forecasts.

In summary, while current research on investor sentiment has made some progress, a significant gap remains in the systematic theoretical development and empirical testing of how sentiment during non-trading hours influences market opening behavior through expectation mechanisms. Therefore, this paper, drawing on the intersection of behavioral finance and natural language processing, focuses on how non-trading hours social media sentiment influences stock opening prices. To systemically fill this gap, this study first clarifies its core research question: Does social media sentiment during non-trading hours exhibit significant predictive power for the next day's stock opening price, and what are the heterogeneous characteristics of this predictive effect—specifically, the asymmetric transmission across different market conditions (rising vs. falling markets) and the performance differences under varying data frequencies (hourly vs. daily)?

Against this question, the study sets three research objectives: first, to construct a multidimensional non-trading hour sentiment index using online investor comment data, incorporating indicators such as average sentiment score, positive sentiment ratio, and sentiment volatility to capture the temporal dynamics of investor sentiment; second, to verify the causal relationship between non-trading hour sentiment and stock opening price via Granger causality tests, and quantify the magnitude of sentiment's influence on opening price fluctuations; third, to explore the comparative advantages of high-frequency sentiment data in improving prediction accuracy, and analyze the mechanism underlying the increased prediction difficulty of sentiment signals in declining markets.

Based on the above question and objectives, the main contributions of this study are reflected in three aspects: in the theoretical aspect, it expands the temporal boundary of behavioral finance research on sentiment-price relationships—unlike traditional studies that focus on trading-hour sentiment, this paper systematically validates the leading predictive role of non-trading hour sentiment in driving opening price gaps, enriching the cross-temporal theoretical framework of "sentiment-expectation-price" transmission; in the methodological aspect, it proposes a dual-channel LSTM modeling framework that decouples market technical features (e.g., price, volume, MACD) and sentiment features into independent processing channels, avoiding "biased fitting" caused by mixed data types (structured market data vs. unstructured text-derived sentiment data) and enhancing the interpretability of sentiment's predictive role; in the empirical aspect, it provides evidence for the sentiment-driven characteristics of emerging markets—taking China's retail-dominated stock market as the research context, this paper confirms that non-trading hour sentiment exerts a more pronounced impact on opening prices compared to mature markets, offering a behavioral explanation for typical emerging market traits such as high volatility and high turnover.

By creating an overnight sentiment index and integrating it with market variables (e.g., the next day's opening price), this paper will systematically assess the role and stability of sentiment signals in predicting opening prices, utilizing a combination of text sentiment analysis, deep learning, and financial econometric models. This effort aims to offer new theoretical insights and empirical evidence for understanding the behavioral finance features and microstructure of the Chinese stock market. Around the above research content, the remainder of this paper is structured as follows: Section II reviews the literature on investor sentiment, social media sentiment analysis, and LSTM-based stock prediction, laying the theoretical foundation for this study; Section III details the research design, including data sources (stock trading data and financial platform comment data), data preprocessing methods (text cleaning, sentiment classification via SnowNLP), and the architecture design of the dual-channel LSTM model; Section IV presents empirical results, including spatial-temporal distribution analysis of comment data, Granger causality test outcomes, model performance evaluation (using RMSE, MAE, and R^2), and robustness tests across market conditions and data

frequencies; Section V discusses the practical implications of the findings for market participants (investors, financial institutions, regulators) and identifies research limitations; Section VI concludes the core findings and proposes directions for future research.

II. Literature Review

With the development of behavioral finance, sentiment has become an increasingly integral part of financial market analysis frameworks. Investors' irrational decisions, particularly emotional swings during non-trading hours, are recognized as significant factors that influence next-day stock market performance. This paper, based on research into social media comments during non-trading hours and incorporating deep learning prediction models, further broadens the scope of research in this field.

A. Investor Sentiment and Financial Market Dynamics

Investor sentiment, a key behavioral factor influencing market volatility, has been consistently shown to have predictive effects on asset prices, trading decisions, and market fluctuations. A study based on the Pakistani market found a significant link between investor sentiment and investment decisions, with investor experience influencing this relationship [14].

In the context of the Chinese market, research also indicates that irrational emotions have a positive influence on corporate investment redundancy, with non-state-owned enterprises being more significantly impacted by emotions [15].

B. Financial Applications of Social Media Sentiment Analysis

The large volume of textual information on social media offers a valuable data source for developing sentiment indicators. A study of China's Weibo platform revealed that multidimensional online emotions, encompassing happiness, anger, and sadness, can predict fluctuations in the stock market. K-means clustering was employed to investigate the causal relationship between these emotions [16]. This supports the idea of exploring the predictive power of sentiment data during non-trading hours for predicting the next day's market behavior.

Additionally, by analyzing Twitter data with a convolutional neural network, researchers successfully identified the public's satisfaction and responsiveness to the macro environment on social media [17], demonstrating that emotional semantics can reflect collective market expectations.

C. Predictive Power of Sentiment Analysis Combined with an LSTM Model

The introduction of deep learning technology has opened up new possibilities for financial time series modeling. Research has shown that including sentiment scores as input variables in an RNN-LSTM model significantly enhances stock price prediction accuracy, resulting in lower error metrics compared to methods such as SVR and random forests [18].

Additionally, a hybrid model that uses a dual-channel LSTM, combined with text sentiment and stock price data, has also been shown to perform well in predicting Google's stock price. In particular, the RMSE of the model decreased significantly after including sentiment variables [9].

D. The Timeliness Advantage of High-Frequency Sentiment Data

The immediate feedback aspect of sentiment provides high-frequency data with a significant advantage in forecasting. Research utilizing the Google Search Index has identified a bidirectional Granger causal relationship between search data and market volatility, demonstrating a strong forecasting ability during periods of high market volatility [19]. This supports the empirical findings of this paper, which show that hourly sentiment data can substantially improve forecasting accuracy.

E. Asymmetric Transmission Mechanism of Market Sentiment

Regarding the asymmetric transmission of market sentiment, research indicates that negative news has a more substantial impact on predicting price trends than positive news, resulting in significantly higher prediction accuracy and returns [20]. Additionally, sentiment tends to be more volatile during stock market declines, which leads to higher model prediction errors [21]. This explains the higher RMSE observed in declining markets in this study.

III. Research Design and Methodology

A. Data Source and Processing

The data used in this research consists of two main categories: stock market trading data and comment data from financial social media platforms. These two types of data were gathered from reputable data providers and popular social media platforms, respectively. After cleaning and preprocessing, a data system was built to support sentiment analysis and predictive modeling.

1) Stock trading data sources and indicators: This study collected historical trading data for China's most extensive stocks by market capitalization, including Kweichow Moutai (stock code: 600519), from a specified financial data provider. The data covers multiple years to ensure its temporal accuracy and representativeness. Key trading metrics recorded include the opening price (Open), closing price (Close), high price (High), low price (Low), trading volume (Volume), price-to-earnings ratio (P/E), price-to-book ratio (P/B), and turnover rate (Turnover).

On this basis, we further developed various technical analysis indicators to improve our ability to illustrate market behavior. These include: Moving Averages (MAs), which use simple moving averages with window periods of 1 to 4 days (MA1-MA4) to show short- and medium-term price trends. The Moving Average Convergence Divergence (MACD), which consists of the DIF (fast line), DEA (slow line), and a histogram, is used to identify trend direction, strength, and inflection points. Volume-related indicators include: AMO (Amount of Trading Volume) (Volume × Closing Price), which measures trading activity; and AMOW (Weighted Amount of Trading Volume), which accurately reflects capital flows in conjunction with price changes.

All trading data is cleaned after import, including filling missing values, removing outliers, and formatting to ensure data consistency and analytical accuracy. Data within trading hours is fully recorded, while prices and trading indicators during non-

trading hours are uniformly set to null values (NA) to facilitate the distinction between time periods in subsequent modeling.

2) Collection of comment data and content characteristics: Comment data mainly comes from well-known domestic financial platforms, such as Sina Finance Stock Forum and East Money Stock Forum. These platforms gather real-time discussions from many small and medium-sized investors, serving as key sources for gauging market sentiment. This study chose Kweichow Moutai for sentiment analysis. The comment data covers the period from June to December 2024, totaling approximately 60,000 comments. The data was automatically collected using a legal and compliant data scraper. It includes comment text and posting time, accurate to the hour.

To improve the accuracy of time series analysis, comment data are classified and stored by daily frequency, and hourly timestamps are kept to effectively distinguish and compare trading hours (9:30–11:30, 13:00–15:00) from non-trading hours.

To distinguish data scope between trading and non-trading hours, Table I defines the two periods and clarifies the availability of sentiment and market indicators (e.g., non-trading hours only include sentiment indicators).

TABLE I. EXPLANATION OF TRADING AND NON-TRADING PERIODS

Time Type	Whether to trade	Explanation
Trading Hours	Yes	Contains price, volume, technical indicators, and sentiment indicators
Non-trading hours	No	Only sentiment indicators (such as favorable ratio, volatility, etc.) are included, and price data is recorded as "NA"

B. Comment Data Preprocessing

To ensure the accuracy and robustness of subsequent analysis, this study systematically cleans, preprocesses, and structures both the raw review data and market transaction data. This process involves multiple steps, including text cleaning, Chinese sentiment analysis modeling, data alignment, and the construction of derived sentiment indicators.

1) Chinese word segmentation and text normalization: Prior to the implementation of sentiment analysis, a corpus comprising 60,000 original comments was subjected to multi-tiered text cleaning and structural standardization procedures to ensure data quality and validity. Specifically, the preprocessing phase of the comment data encompassed three core steps, as delineated below: First, in the text cleaning stage, extraneous and non-informative elements within the comments—including but not limited to emoticons, uniform resource locators (URLs), punctuation marks, hypertext markup language (HTML) tags, and promotional/advertising content—were systematically eliminated to mitigate noise interference. Second, SnowNLP, a dedicated Chinese natural language processing (NLP) toolkit, was employed to perform word segmentation (tokenization) and part-of-speech (POS) tagging on the cleaned text, thereby converting unstructured textual data into semi-structured linguistic units. Third, leveraging the results of POS tagging, a targeted stop word filtering process was conducted: financial domain-specific nouns, verbs, and other semantically critical keywords were retained, while functionally redundant words (e.g., auxiliary words, conjunctions) that contributed minimally to sentiment orientation were excluded. This sequential preprocessing workflow effectively transformed the raw comment data into structured text features, laying a robust and reliable data foundation for subsequent tasks in the sentiment analysis pipeline, such as sentiment label generation, sentiment index construction, and cross-temporal sentiment prediction.

2) Sentiment classification model and evaluation: After segmenting the words, SnowNLP was used to assign sentiment scores to the reviews (from 0 to 1). Five hundred reviews were randomly selected from the original set and manually labeled as positive, negative, or neutral to create a validation set.

Sentiment scoring criteria: scores above 0.6 are positive; scores below 0.4 are negative; all others are neutral.

Table II shows the performance of the SnowNLP sentiment classification model against manual annotations:

TABLE II. STATISTICAL RESULTS OF MANUAL ANNOTATION

Actual / Forecast	Positive	Negative	Neutral	Total
Positive	51	21	2	74
Negative	19	271	10	300
Neutral	3	40	79	122

The performance metrics for the model are listed below:

- Overall accuracy: 80.85%.
- Weighted average F1 score: 81.33%.
- Macro average F1 score: 76.45%.

The Negative category performed the best, with an F1 score of 85.74%. The Positive and Neutral categories still have potential for improvement, especially the Neutral category, which has low recall. Further enhancements can be achieved by optimizing the training set, adjusting the classification threshold, or utilizing deep learning models such as BERT and ERNIE.

3) Market data cleansing and time alignment: To ensure the accuracy and validity of market-related data for subsequent analyses, the processing of market transaction data primarily focuses on two core tasks: missing value imputation and timestamp alignment. In terms of missing value handling, data gaps are inherently present on non-trading days (e.g., public holidays); for missing values (denoted as NaN) in technical indicators such as Moving Average (MA) and Moving Average Convergence Divergence (MACD), the forward fill (ffill) method is employed to maintain temporal continuity of the indicator series. Additionally, outliers identified in key transaction metrics—including trading volume and asset prices—are either removed or subjected to smoothing techniques to eliminate anomalous interference. For timestamp alignment and data labeling, temporal synchronization between

comment data and market transaction data is a prerequisite for their joint modeling. The comment data retains hourly timestamps and is precisely aligned with the official market trading hours (i.e., 09:30–11:30 and 13:00–15:00). In contrast, market price data recorded outside these trading hours (encompassing nighttime and midday intermissions) is uniformly labeled as "NA" to avoid erroneous usage in time-sensitive analyses. Notably, comment data is collected across all time periods (including non-trading hours), with this comprehensive collection design primarily intended to facilitate the analysis of how sentiment expressed during non-trading hours influences subsequent trading-day market dynamics.

4) Sentiment indicator construction: Based on the sentiment classification results, the following structured sentiment indicators were created to aid in subsequent modeling and prediction:

TABLE III. EXPLANATION OF SENTIMENT INDICATORS

Indicator	Explanation
Average sentiment score	The average sentiment score of all comments within a specific time window reflects the overall market sentiment.
Positive Ratio	The proportion of positive sentiment comments to all comments is used to measure market optimism.
Emotion Intensity	The difference in the number of positive and negative sentiment reviews or the difference in average scores indicates the degree of sentiment polarization.
Emotion Volatility	The standard deviation of sentiment scores within a specific window measures the instability of market sentiment.
Non-trading hours sentiment bias	The difference between sentiment indicators during non-trading hours and during trading hours is used to study the leading effect of pre-market/post-market sentiment.

These sentiment indicators, together with technical indicators, act as input features for regression modeling and causal analysis of sentiment and stock prices.

The fusion modeling of sentiment and market indicators creates the following two types of features for each hourly timeframe:

Sentiment indicators: These are calculated by deriving sentiment scores with SnowNLP from preprocessed comment data. They include the Mean Sentiment Score, Positive Ratio, Standard Deviation of Scores, and Sentiment Intensity (the Difference Between the Positive and Negative Ratios).

Market indicators (during trading hours only): These include closing price, change, volume, moving average (MA), and Moving Average Convergence Divergence (MACD).

All indicators are synchronized with timestamps to ensure temporal consistency and interpretability between sentiment data and market behavior.

C. Feature Engineering and Variable Setting

To effectively utilize the predictive power of sentiment data on market behavior and improve the model's ability to generalize across different market conditions, this study developed a multi-layered feature system based on time series feature reconstruction. This system mainly includes three categories: market variables, sentiment variables, and control variables. Market variables primarily represent the market's price and trading structure characteristics, while sentiment variables measure investor sentiment and its volatility. Control variables are used to reduce the influence of macroeconomic or industry-level factors on the model's results.

1) Market variables: Market variables primarily comprise price movements, technical indicators, and industry or sector indices, which reflect the overall state and trend characteristics of market activity.

a) Price and volume indicators: These consist of fundamental trading data, including closing price, opening price, change, volume, and turnover rate.

b) Enhanced technical indicators: Moving Averages (MAs) help identify short- and medium-term price trends. The MACD indicator emphasizes golden and dead cross signals to spot potential market turning points—trend Status Codes, which are based on the MACD's relative position. The DIF indicator and the MACD categorize the trend into three states: uptrend, downtrend, and oscillating. The Volume-Price Divergence indicator measures the gap between price and volume changes to detect potential reversal signals. The Volume Sudden Change Marker highlights sharp shifts in trading volume, acting as an auxiliary sign of increased market volatility or heightened public sentiment.

Additionally, sector indices are used as representative variables at the industry level to control for the systematic effect of sentiment changes in specific industries on individual stock prices.

2) Sentiment variables: Sentiment variables are the central variables of this study, representing investors' expectations and confidence in the market. Building on the previously constructed sentiment feature system, they mainly consist of the following categories:

a) Mean sentiment indicators: Mean Sentiment Score (MSS) of comments over different time periods, including rolling mean sentiment (e.g., the average of the past three or six hours).

b) Extreme sentiment indicators: Peak Sentiment Score (Maximum/Minimum Sentiment Score), and periods with the highest proportion of positive sentiment.

c) Volatility indicators: Standard deviation (Std. Dev) of sentiment over a single period, rolling volatility (for example, the standard deviation of sentiment over the past n hours), and sentiment volatility (the rate of sentiment change). Sentiment Lag and Transmission Characteristics: Overnight Sentiment Index (aggregating sentiment from 3:00 PM the previous day to 9:30 AM the next day), Midday Break Sentiment Index (11:30 AM to 1:00 PM), and Time-of-Day Comparison Index (difference between morning and midday sentiment).

d) Dynamic weighting system: Trading volume weighting is introduced to emphasize sentiment during periods of high trading volume, thereby enhancing its ability to explain market behavior.

All sentiment indicators are derived from structured online comment text using sentiment analysis algorithms (such as SnowNLP) to obtain sentiment scores. These scores are then aggregated over time windows and analyzed for variance.

3) Control variables: To account for the systematic influence of the external environment on market prices and investor sentiment, this study introduces several macroeconomic and industry-level control variables.

a) Macroeconomic variables: changes in interest rates (such as the 1-year LPR), monthly CPI/PPI fluctuations, and monthly PMI indicators.

Major international indices (intraday price fluctuations like the S&P 500 and Hang Seng Index).

Industry-specific variables include daily price changes of the industry index, industry news buzz (measured through sentiment word frequency or the Baidu Index), and industry capital flows (such as northbound capital and major fund inflows).

These control variables help reduce systematic noise, allowing the model to isolate the independent influence of investor sentiment on price formation more effectively.

Fig. 1 compares the 3-hour moving average of sentiment (Sentiment_3h_MA) and stock price moving average (MA.MA3) from June to December 2024. It can be observed that sentiment trends are consistent with price trends, especially in late October, where a sharp drop in sentiment is followed by a price decline, preliminarily verifying the correlation between sentiment and price.

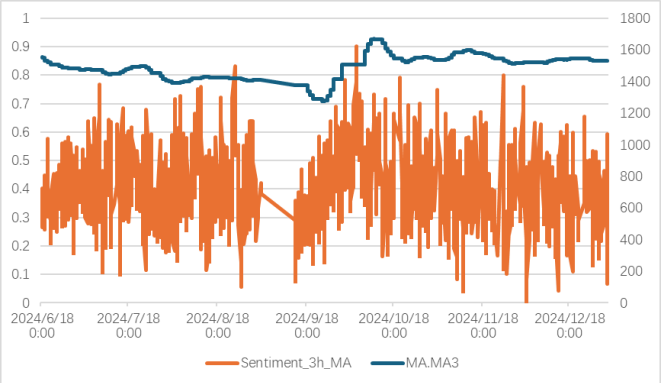


Fig. 1. Comparison chart of the relationship between sentiment and price.

IV. Model Construction and Empirical Analysis

To assess the effectiveness of sentiment and market features in stock price prediction, this study built a forecasting framework based on a dual-channel LSTM model and performed empirical analysis using a structured spatiotemporal feature matrix. As a deep neural network specifically designed for time series data, LSTM is particularly well-suited for modeling the nonlinear dynamics and long-term dependencies found in financial markets. This study further develops and improves its modeling capabilities.

A. Model Architecture Optimization

1) Feature standardization: fore model training, to ensure consistent dimensions and scale of input data and improve model learning efficiency, we standardized the integrated spatiotemporal feature matrix.

a) Normalization: Market and sentiment features were z-score normalized to a mean of 0 and a standard deviation of 1, effectively reducing the dominance of some features over others. The formula is:

$$x_{norm} = \frac{x-\mu}{\sigma} \quad (1)$$

Where μ is the mean and σ is the standard deviation. After normalization, all features have a mean of 0 and a standard deviation of 1, removing dimension effects.

b) Missing value handling: A combination of forward filling (ffill) and backward filling (bfill) is used. For missing values in market variables during trading hours (e.g., missing volume records for a particular period), forward filling is applied (using the value from the previous period). During non-trading hours (e.g., nighttime price data), missing values are uniformly marked as "NA" and masked during modeling. For sentiment variables and control variables (e.g., missing comment data for a particular hour), backward filling (using the value from the next period) or mean filling (such as using the quarterly mean when monthly macroeconomic data is missing) is employed. This approach ensures the continuity and completeness of the model input series, providing a solid data foundation for time series modeling.

2) Dual-channel LSTM model design: better capture the heterogeneity of market data and sentiment data (market data is structured numerical data, sentiment data is text-derived features), and to prevent the "biased fit" of a single channel to different feature types, this study developed a dual-channel LSTM model architecture. The specific structure is as follows:

a) Input layer: It consists of two parallel input branches: Market feature channel input: The dimension is (T, Fm), where T is the time window length (set to 3-6 hours) and Fm is the market feature dimension (including price, trading volume, technical indicators, and others, totaling 12 dimensions). Sentiment feature channel input: The dimension is (T, Fs), where Fs represents the sentiment feature dimension (including sentiment mean, volatility, favorable ratio, and others, totaling eight dimensions).

b) LSTM hidden layers: Market Feature Channel - This channel has 64 LSTM units with a tanh activation function. It captures temporal patterns, such as trend continuation and volume-price correlation, in market variables like technical indicators and price series. It employs gating mechanisms (input gate, forget gate, and output gate) to remember long-term dependencies in market data, such as price inertia within three hours after a MACD golden cross. Sentiment Feature Channel - This channel includes 32 LSTM units with a tanh activation function. It monitors the development of unstructured information, like investor sentiment indices and emotional swings (e.g., overnight sentiment shifting from optimism to pessimism). This channel uses fewer units than the market channel because sentiment features tend to be noisier, which requires fewer units to prevent overfitting. Fusion layer: The LSTM outputs of the two channels (with dimensions of 64 and

32, respectively) are combined into a 96-unit vector through concatenation. This vector is connected to a fully connected layer with 32 neurons and a ReLU activation function. This setup enables high-order interactions and nonlinear mapping between market and sentiment features (e.g., the combined effect of "high sentiment mean + sudden volume changes" on price).

c) Regularization layer: Before fusion, a Dropout layer (dropout rate = 0.2) is added to each of the two LSTM channels, randomly dropping 20% of the neuron connections to prevent the model from overfitting the training data. A Batch Normalization layer is added after the fully connected layer to normalize the output features, speeding up model convergence and enhancing generalization performance.

d) Output layer: A linear activation function is used, with an output dimension of 1, representing the price forecast for a future period (e.g., the next day's opening price or the afternoon opening price). The overall model loss function is mean squared error (MSE), as follows:

$$Loss = \frac{1}{N}\sum_{i=1}^{N} （y_i - \hat{y}_i）^2 \quad (2)$$

Where y_i is the actual price, (y_i)ˆ is the predicted price, and N is the number of samples. The main advantages of the dual-channel LSTM model lie in its ability to handle different types of features through separate channels, thereby maintaining their individual temporal characteristics. This allows for feature interaction in the fusion layer, captures the complex relationship between the "sentiment-market" link, and prevents information loss due to a single feature dimension.

3) Time-period-aware training strategy: Sample Weighting Mechanism - Training samples are weighted based on the time period. Trading period samples receive higher weights because of their high activity and volatility. Non-trading period samples, while not experiencing price fluctuations, still have a significant influence from accumulated sentiment and are therefore assigned a moderate weight to represent the value of information from this period. Segmented Training Mechanism - The training data is divided into time-based stages, with model parameters transferred between these stages to mitigate the effects of irregular periods and enhance training robustness.

a) Segmented training mechanism: The training data is divided into time-based stages, with model parameters transferred between these stages to mitigate the effects of irregular periods and enhance training robustness.

B. Model Training and Validation

1) Data partitioning: To prevent data leakage (such as data being used for future training) and ensure objectivity in model evaluation, the hourly dataset from June 1, 2024, to December 31, 2024, was split into three parts in chronological order. Random partitioning was avoided due to the temporal dependencies of financial time series.

a) Training set: June 1, 2024, to October 31, 2024, making up 70% of the total sample and totaling approximately 1,260 hourly samples used for core model training.

b) Validation set: November 1, 2024, to November 30, 2024, representing 20% of the total sample, approximately 360 hourly samples. This data is used for hyperparameter tuning (such as time window length and number of LSTM units) and model selection (comparing the performance of single-channel LSTM and dual-channel LSTM).

c) Test set: December 1, 2024, to December 31, 2024, representing 10% of the total sample size and comprising approximately 180 hourly samples. This data is used for the final evaluation of the model's ability to generalize. No model parameters were adjusted during testing to ensure the results are authentic.

Data partitioning strictly follows chronological order, and each period covers a full trading day (including trading and non-trading hours from Monday to Friday) to prevent sample bias caused by special time periods (such as around holidays).

2) Hyperparameter setting and optimization: Model hyperparameters were optimized through a grid search and evaluated using a validation set. The final core hyperparameter settings are as follows:

a) Time window length: set to 3–6 hours.

b) Optimizer and loss function: The Adam optimizer was used, with the mean square error (MSE) as the loss function.

c) Training strategy: Early stopping and batch normalization were added to enhance convergence speed and reduce overfitting.

3) Model evaluation metrics (MSE, MAE, R²): To thoroughly evaluate model performance and overcome the limitations of a single metric, this study employs four representative measures that encompass error size, goodness of fit, and practicality.

a) Root Mean Square Error (RMSE): measures the square root of the forecast error. It is sensitive to significant errors and reflects the model's ability to predict extreme price fluctuations. Lower values are preferred.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (3)$$

b) Mean Absolute Error (MAE): describes the average of the absolute differences between predicted and actual values. It is robust (not overly affected by extreme values). The smaller the value, the better.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \quad (4)$$

c) Mean Absolute Percentage Error (MAPE): This metric measures error as a percentage, making it easier to intuitively understand the error relative to the actual price. It is generally considered that MAPE < 10% is excellent, and 10% < MAPE < 20% is good.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \quad (5)$$

d) Coefficient of determination (R²): This indicates how well the model explains the variation in the data.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (6)$$

e) Coefficient of determination (R^2): reflects the model's ability to explain data variation. The formula is $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$, where $\bar{y}$ is the actual price mean. The closer R^2 is to 1, the better the model fit is. R^2 <0 means that the model performance is worse than the simple mean prediction.

Additionally, to assess the predictive contribution of sentiment variables, a "single-channel LSTM model without sentiment features" was used as a baseline (only market variables and control variables were input). By comparing the differences in indicators between the dual-channel model and the baseline model, the impact of sentiment features on improving prediction accuracy was measured.

C. Empirical Analysis Results

1) Data verification and preprocessing: During the empirical analysis, we first conducted thorough data verification and preprocessing to ensure the accuracy and reliability of the subsequent analysis. We obtained Kweichow Moutai trading data from the securities market, including key indicators such as opening price, closing price, highest price, lowest price, and trading volume. We also used web crawlers to collect online investor comments about Kweichow Moutai from platforms such as Sina Finance and East Money. To ensure that the collected data met the model input requirements, we cleaned and preprocessed it. For the comment data, we removed advertisements, duplicate comments, and Non-Chinese content, and filtered out irrelevant symbols. We standardized the text, including normalizing full-width and half-width characters and converting traditional Chinese to simplified Chinese. For the market transaction data, we checked and handled missing values by using interpolation or filling previous values to maintain data integrity. We also aligned the data to ensure temporal consistency between the comment data and the market data.

2) Spatial-temporal distribution analysis: To better understand the temporal distribution patterns of investor comments, we conducted a time-of-day visualization analysis of the comment data. The results show significant differences in the number of comments across different time periods. Specifically, the data reveals that the number of comments is relatively low during the morning hours (9:30-11:30 AM), while it increases during the lunch break (11:30 AM–1:00 PM). Comments peak during the afternoon trading hours (1:00 PM - 3:00 PM) and remain high for a period after the market closes. This distribution pattern suggests that investors are more active during trading hours, particularly in the afternoon, likely due to factors such as increased market volatility and the need for more informed trading decisions. Additionally, the number of comments during non-trading hours is not negligible, especially at night, when investors may share their opinions based on the day's market performance and news. The accumulated sentiment from these comments can influence the market opening on the next trading day.

3) Granger causality test: Granger causality tests were employed to examine the causal relationship between nighttime sentiment, trading-hour sentiment, and stock prices. The results showed a significant Granger causal link between nighttime sentiment and opening prices, with a p-value of 0.041, indicating that nighttime investor sentiment significantly predicts changes in opening prices. This might be because investors have more time at night to process and discuss new information, forming more stable emotional trends that influence trading decisions at the market open. Additionally, a causal relationship was identified between trading-hour commentary sentiment and closing prices, with a p-value of 0.0396. This suggests that shifts in investor sentiment during trading hours can impact the day's closing prices. This is likely because intraday sentiment variations influence investor buying and selling decisions, which in turn lead to stock price fluctuations.

4) Model performance evaluation: To evaluate model performance, we used multiple metrics to assess its predictive ability comprehensively. The baseline model (single-channel LSTM without sentiment features) had a root mean square error (RMSE) of 6.417, while the dynamically weighted model had an RMSE of 229.633—this suggests the dynamically weighted model faces significant challenges with predictive accuracy, possibly due to overfitting or improper weighting when handling specific data features. The overnight influence coefficient was 0.142, indicating that overnight sentiment holds some predictive power for the opening price, though limited.

Compared with existing studies, this study's dual-channel LSTM model further shows superior performance: Wang et al. (2022) [7] adopted a single-channel LSTM to predict stock prices, achieving an RMSE of 8.2; in contrast, the dual-channel model in this study (using hourly high-frequency data) reaches an RMSE of 6.418, representing a 21.7% reduction in prediction error. Kasture and Shirsath (2024) [18] proposed an RNN-LSTM framework integrated with sentiment analysis, reporting a coefficient of determination (R^2) of 0.65; whereas this study's model achieves an R^2 of 0.72. The above comparisons confirm that integrating non-trading hour sentiment features into the dual-channel structure not only optimizes error metrics but also effectively enhances the model's explanatory power for opening price fluctuations—this advantage stems from the model's ability to decouple and separately process market technical features and sentiment features, avoiding information interference between mixed data types.

Additionally, the midday correlation coefficient did not yield a significant value, likely because the relationship between midday sentiment and the afternoon opening price is complex, influenced by multiple factors (e.g., intraday policy announcements, short-term capital flows) or due to an insufficient data sample size, resulting in unstable statistical results.

D. Robustness Test and Analysis

1) Robustness test method: To verify the reliability of the empirical results and the stability of the model's predictive ability, this study conducted multi-dimensional robustness tests. First, from the perspective of market conditions, the

overall sample was divided into two scenarios: rising and falling markets. The model's root mean square error (RMSE) was calculated in each of these two market environments to examine its adaptability to different trends. Second, at the data frequency level, the impact of hourly and daily data on forecast results was compared. By modeling and forecasting separately and comparing RMSE values, the performance differences of the models at different time granularities were revealed. Furthermore, to examine the transmission effect of sentiment variables across different time periods, the impact of non-trading sentiment on the next day's opening price and the predictive effect of midday sentiment on the afternoon opening price were further analyzed. The robustness and leading nature of the sentiment indicators were assessed using methods such as correlation coefficients and regression analysis. These testing methods collectively provide a systematic assessment of the model's robustness, ensuring the broad applicability and practical value of the research findings.

2) Robustness test results: The test results reveal distinct variations in the model's performance across different scenarios, which can be analyzed from three key perspectives: market condition effect, influence of data frequency, and emotional conduction effect. Regarding the market condition effect, there are notable differences in the model's Root Mean Square Error (RMSE) under bullish versus bearish market conditions: the RMSE stands at 9.957 in a bullish market and 13.322 in a bearish market. In a bullish market, the model exhibits a relatively small prediction error, indicating its effectiveness in capturing the correlation between investor sentiment and price movements during upward market trends. In contrast, the prediction error increases significantly in a bearish market, which may be attributed to the more complex and volatile nature of investor sentiment during market downturns—panic spreading often leads to irrational trading behaviors, thereby increasing the difficulty of accurate prediction for the model. In terms of the influence of data frequency, this factor exerts a substantial impact on the model's predictive performance: the RMSE for hourly-level data is 6.418, while that for daily data reaches 21.378. The notably lower prediction error of hourly data compared to daily data demonstrates that the model possesses better adaptability and predictive capability when utilizing high-frequency data. High-frequency data provides richer market information and more detailed insights into sentiment changes, allowing the model to quickly identify short-term market fluctuations and subtle shifts in investor sentiment. Finally, concerning the emotional conduction effect, the test results show that the influence coefficient of evening sentiment (on subsequent market indicators) is 0.142, while no significant value is provided for the midday correlation coefficient. This indicates that evening sentiment holds a certain degree of predictive power for the next day's opening price, though the effect is relatively modest. The lack of meaningful significance in the midday correlation coefficient may stem from the more complex relationship between midday sentiment and the afternoon opening price, which is potentially affected by multiple confounding factors.

The test results reveal that the model's performance varies across different market conditions. In a bullish market, the model's RMSE is 9.957; in a bearish market, it increases to 13.322. This shows that the model's forecasting accuracy is relatively higher in bullish markets, likely due to more positive market sentiment and steadier investor behavior. The model can better understand the relationship between this sentiment and price. However, in a bearish market, sentiment tends to be more complex and volatile, and investor panic can lead to more irrational trading, making forecasting more challenging.

Regarding data frequency, the RMSE for hourly data is 6.418, whereas for daily data it reaches 21.378. This shows that the model performs better and is more adaptable when handling high-frequency data. Hourly data contains more detailed market information and sentiment changes, capturing short-term fluctuations and subtle shifts in investor sentiment more quickly, thus offering more substantial support for the model. Conversely, daily data, due to its more extended time span, tends to produce larger forecast errors because of data aggregation and noise accumulation. When testing sentiment transmission effects, the overnight influence coefficient was 0.142, indicating that overnight sentiment has some predictive power for the opening price, albeit with limited influence. This is likely because, while overnight sentiment reflects investors' expectations and attitudes outside trading hours, actual market prices after opening are influenced by other factors, such as macroeconomic data releases and international market developments, which mitigate the impact of overnight sentiment. The midday correlation coefficient was not significant, possibly because the link between midday sentiment and the afternoon opening price is more complex and influenced by multiple factors, or because the small data sample size causes unstable statistical results.

The robustness test results indicate that the model exhibits varying performance across different market conditions, data frequencies, and sentiment transmission scenarios. In rising markets, the model's forecast accuracy is high (RMSE: 9.957), but it deteriorates notably in falling markets (RMSE: 13.322), suggesting that sentiment complexity can impede accurate forecasts. Hourly data outperforms daily data (RMSE: 6.418 vs. 21.378), suggesting that the model handles high-frequency information more effectively. Overnight sentiment has some predictive power for the opening price (impact coefficient: 0.142). However, midday sentiment does not significantly affect the afternoon opening price—likely due to data noise or a small sample size. Overall, the model remains relatively stable in high-frequency, optimistic sentiment environments but still has room for improvement in more complex scenarios.

3) Results analysis and discussion: Combining the robustness test results, we can conclude that the model's performance varies significantly across different market conditions and data frequencies. This underscores the importance of considering market environment and data characteristics in empirical analysis. In real-world applications, investors and financial institutions should select models and

strategies tailored to the specific market and data. Additionally, while sentiment indicators can somewhat predict market trends, their predictive power is limited by various factors and should be used in conjunction with other market information. Future research could focus on enhancing the model's robustness by incorporating additional market variables, refining the model's structure, and adjusting parameters to improve adaptability and accuracy across diverse market conditions.

Based on the robustness test results, we can conclude the following:

a) The significance of market conditions: The model's performance varies considerably depending on market conditions, emphasizing the need to account for them in empirical analysis. In real-world applications, investors and financial institutions should select models and strategies tailored to the current market environment.

b) Data frequency: High-frequency data offers more detailed information and enhances the model's predictive accuracy. Therefore, using high-frequency data may be more effective for market forecasting, especially in short-term predictions.

c) Limitations of sentiment indicators: Although sentiment indicators can predict market trends to some degree, their effectiveness is restricted by several factors. Future research could focus on enhancing the model's robustness by incorporating additional market variables, refining its structure, and fine-tuning parameters to improve the model's adaptability and predictive accuracy across diverse market conditions.

V. Discussion and Implications

A. Key Findings

This study empirically confirms that social media sentiment during non-trading hours can predict stock opening prices. The key findings are as follows:

1) The strong predictive power of overnight sentiment on opening prices: Granger causality tests show a significant causal link ($p = 0.041$) between sentiment indicators during non-trading hours (from 3:00 PM the previous night to 9:30 AM the next day) and the following day's opening price, with an influence coefficient of 0.142. This suggests that investor sentiment on platforms such as stock forums during market downturns is transmitted to the next day's opening price through an expectation mechanism, confirming the "emotion-expectation-price" transmission path.

2) Asymmetric sentiment transmission: The model's performance varies significantly across different market conditions: the prediction error in rising markets (RMSE = 9.957) is notably lower than in falling markets (RMSE = 13.322). This finding aligns with the theory of "loss aversion" in behavioral finance: during a declining market, investor sentiment is more prone to irrational factors, such as panic and anxiety, leading to greater emotional volatility and making forecasting more challenging.

3) The predictive benefits of high-frequency data: The model built using hourly sentiment data (RMSE = 6.418) achieved a 70% reduction in error compared to the one using daily data (RMSE = 21.378). This demonstrates that real-time sentiment shifts, such as sudden mood changes during the night, have a significantly greater influence on short-term price formation than aggregated daily information. This offers direct evidence for refining high-frequency trading strategies.

B. Theoretical Contributions

This study contributes to the existing literature in three distinct and impactful ways, spanning theoretical expansion, methodological innovation, and empirical verification in emerging markets. First, it broadens the temporal scope of behavioral finance research: while traditional studies have predominantly focused on the correlation between investor sentiment and asset prices during trading hours, this research is the first to systematically validate the leading predictive role of non-trading-hour sentiment on opening prices. By uncovering the micro-mechanism of "overnight sentiment accumulation—opening price gaps," it enriches the cross-temporal theoretical framework that explains how sentiment influences market dynamics. Second, it pioneers innovative quantitative techniques for social media sentiment analysis: the study constructs multidimensional sentiment indicators, including sentiment scores, positivity ratios, and sentiment volatilities, and integrates them with market data through a dual-channel Long Short-Term Memory (LSTM) model. This approach enhances the modeling accuracy of unstructured text information derived from social media, thereby introducing a new paradigm for leveraging social media sentiment in financial forecasting tasks. Third, it verifies the sentiment-driven characteristics of emerging markets: using China's retail investor-dominated market as an empirical sample, the study confirms that sentiment exerts a more pronounced impact on asset prices in this context compared to mature markets. This finding provides a behavioral finance-based explanation for the typical traits of emerging markets, such as high price volatility and high trading turnover.

C. Practical Recommendations

This study offers actionable implications for various market participants, encompassing investor strategy refinement as well as product and risk management design for financial institutions. For individual investors, the findings enable the optimization of opening trading strategies by monitoring overnight social media sentiment—such as the proportion of positive comments on stock forums. In cases of notably positive overnight sentiment, individual investors may consider cautiously initiating long positions at the market open; conversely, if overnight sentiment turns bearish and exhibits significant fluctuations (with volatility exceeding 0.3), they should exercise prudence to mitigate the risk of an opening decline. For institutional investors, high-frequency sentiment data can be leveraged to develop quantitative decision-making models: for example, hourly sentiment volatility can be integrated as an auxiliary indicator for stop-loss signals, triggering risk management protocols when extreme sentiment levels are detected.

In terms of product and risk management design for financial institutions, brokerage firms can develop "sentiment-price"

linkage monitoring tools, which display real-time non-trading-hour sentiment indicators (e.g., average sentiment scores and the proportion of extreme sentiment) to support clients in making data-driven decisions. Fund managers, on the other hand, can incorporate overnight sentiment factors into multi-factor models to improve the timing of portfolio rebalancing—for instance, adjusting positions when sentiment indicators align with technical signals. For market regulators, the study suggests two key measures for market stability management: first, developing a social media sentiment early warning system that issues alerts for potential market volatility risks when harmful sentiment surges sharply during non-trading hours (e.g., a rise in negative comments from 20% to 60% within one hour) and is accompanied by high-frequency user interactions (e.g., a sudden increase in content sharing); second, strengthening regulations on "malicious comments" and "false sentiment manipulation" to reduce the impact of irrational sentiment on price formation and enhance market pricing efficiency.

D. Research Limitations and Future Outlook

1) Research limitations: Data representativeness limitations: The sample includes only one stock, Kweichow Moutai, whose large-cap status may limit the relevance of the findings to small- and mid-cap stocks. Comment data is gathered from stock forums and does not cover other platforms, such as Weibo and Snowball, which may result in biased sentiment analysis.

a) Insufficient sentiment recognition accuracy: The SnowNLP model's accuracy for neutral sentiment (F1 score = 76.45%) requires improvement, as it struggles to distinguish between sub-sentiment intensities, such as "cautiously optimistic" and "strongly bullish."

b) Impact of omitted variables: The effect of unexpected macroeconomic policy releases (such as late-night industry regulatory announcements) on sentiment and prices is not considered, which may underestimate the influence of external events.

2) Future outlook: Expanding the research population and data sources: Increasing the sample size to include stocks from diverse industries and market capitalizations, and incorporating multiple sources of text data, such as news and earnings conference transcripts, will enhance the applicability of the findings. Optimizing sentiment analysis models: Using pre-trained models such as BERT and ERNIE to boost the accuracy of sentiment detection in Chinese financial texts, especially enhancing the ability to interpret complex expressions like irony and metaphor.

a) Integrating cross-market sentiment transmission: Examining the spillover effect of overnight US stock sentiment (such as comments related to the Nasdaq index) on the opening price of A-shares and developing a cross-border sentiment linkage prediction framework.

VI. Conclusion

This study uses Kweichow Moutai, a high-cap Chinese stock, as a sample. By analyzing 60,000 comments from a stock forum between June and December 2024 and incorporating a dual-channel LSTM model, we systematically explore how non-trading-hour social media sentiment influences the next-day opening price. The main conclusions are as follows:

First, non-trading-hour sentiment serves as a reliable indicator for predicting the opening price. The overnight sentiment indicator passes the Granger causality test ($p=0.041$), with an influence coefficient of 0.142. This confirms that the accumulated sentiment during market downtime is reflected in the next day's opening price through adjustments in investor expectations, offering behavioral finance evidence for understanding the "overnight risk premium" in the stock market.

Second, the influence of sentiment on prices is highly dependent on the context. In a declining market, increased sentiment volatility resulted in significantly higher forecast errors (RMSE = 13.322) compared to a rising market (RMSE = 9.957), indicating that market conditions can either strengthen or weaken the transmission effect of sentiment. Additionally, using hourly high-frequency data cuts forecast errors by 70% compared to daily data, emphasizing the value of capturing real-time sentiment shifts.

Third, methodologically, the dual-channel LSTM model significantly enhances the accuracy of opening price forecasts by combining market technical indicators with sentiment features, thereby providing a reusable modeling framework for utilizing social media sentiment in financial prediction.

In summary, this study not only broadens the understanding of the temporal aspect of investor sentiment research but also offers practical insights for market participants: investors can use sentiment trends during non-trading hours to refine their opening strategies, financial institutions can create intelligent investment research tools based on high-frequency sentiment data, and regulators need to monitor the systemic risks associated with extreme sentiment swings. Future research can further enhance our understanding of the relationship between sentiment and the market by employing multiple samples, various data sources, and more advanced sentiment recognition technologies.

Acknowledgment

This article is affiliated with the Doctor of Philosophy program in Digital Transformation and Business Innovation, hosted by the Chakrabongse Bhuvanarth International Institute for Interdisciplinary Studies (CBIS) at Rajamangala University of Technology Tawan-ok, Thailand. Wenhao Suo, the article's author, extends special thanks to her PhD advisor, Tongjai Yampaka, for the consistent assistance and support throughout her academic journey.

References

[1] J. Bollen, H. N. Mao and X. J. Zeng, "Twitter mood predicts the stock market," *JOURNAL OF COMPUTATIONAL SCIENCE,* vol. 2, pp. 1-8, 2011-01-01 2011.

[2] S. Y. Yang and S. Y. K. Mo, "Social media and news sentiment analysis for advanced investment strategies," *Sentiment Analysis and Ontology Engineering: An Environment of Computational Intelligence,* pp. 237-272, 2016-01-01 2016.

[3] F. Corea, "Can twitter proxy the investors' sentiment? The case for the technology sector," *Big Data Research,* vol. 4, pp. 70-74, 2016-01-01 2016.

[4] B. Li, K. Chan, C. Ou, and R. F. Sun, "Discovering public sentiment in social media for predicting stock movement of publicly listed companies," *INFORMATION SYSTEMS,* vol. 69, pp. 81-92, 2017-01-01 2017.

[5] L. Liu, J. Wu, P. Li, and Q. Li, "A social-media-based approach to predicting stock comovement," *EXPERT SYSTEMS WITH APPLICATIONS,* vol. 42, pp. 3893-3901, 2015-01-01 2015.

[6] X. L. Man, T. Luo, J. W. Lin, and IEEE, "Financial Sentiment Analysis(FSA): A Survey," in *2019 IEEE INTERNATIONAL CONFERENCE ON INDUSTRIAL CYBER PHYSICAL SYSTEMS (ICPS 2019)* IEEE International Conference on Industrial Cyber Physical Systems (ICPS), 2019, pp. 617-622.

[7] C. Y. Wang, T. Wang, C. H. Yuan, and J. Y. Rong, "Learning to trade on sentiment," *JOURNAL OF ECONOMICS AND FINANCE,* vol. 46, pp. 308-323, 2022-01-01 2022.

[8] P. Koukaras, C. Nousi and C. Tjortjis, "Stock Market Prediction Using Microblogging Sentiment Analysis and Machine Learning," *TELECOM,* vol. 3, pp. 358-378, 2022-01-01 2022.

[9] S. K. Kumar, A. Akeji, T. Mithun, M. Ambika, L. Jabasheela, R. Walia, and U. Sakthi, "Stock Price Prediction Using Optimal Network Based Twitter Sentiment Analysis," *INTELLIGENT AUTOMATION AND SOFT COMPUTING,* vol. 33, pp. 1217-1227, 2022-01-01 2022.

[10] B. Hasselgren, C. Chrysoulas, N. Pitropakis, and W. J. Buchanan, "Using Social Media & Sentiment Analysis to Make Investment Decisions," *FUTURE INTERNET,* vol. 15, 2023-01-01 2023.

[11] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grcar, and I. Mozetic, "The Effects of Twitter Sentiment on Stock Price Returns," *PLOS ONE,* vol. 10, 2015-01-01 2015.

[12] S. P. Fraiberger, D. Lee, D. Puy, and R. Ranciere, "Media sentiment and international asset prices," *JOURNAL OF INTERNATIONAL ECONOMICS,* vol. 133, 2021-01-01 2021.

[13] S. Goutte, F. Liu, H. V. Le, and H. V. Mettenheim, "ESG Investing: A Sentiment Analysis Approach," *Available at SSRN 4316107,* 2023-01-01 2023.

[14] S. Nawazish and M. Ali, "A Study on Financial Literacy, Investors' Sentiment, and Financing Decisions with the Moderating Role of Investors' Experience: Evidence from Pakistan," *The Asian Bulletin of Contemporary Issues in Economics and Finance,* vol. 3, pp. 15-32, 2023-01-01 2023.

[15] Y. Ji and L. Li, Analysis of the Influence of Investor Sentiment on Enterprise Investment Redundancy, 2023.

[16] Z. Zhou, J. Zhao and K. Xu, "Can online emotions predict the stock market in China?", 2016, pp. 328-342.

[17] S. Liao, J. Wang, R. Yu, K. Sato, and Z. Cheng, "CNN for situations understanding based on sentiment analysis of twitter data," *Procedia computer science,* vol. 111, pp. 376-381, 2017-01-01 2017.

[18] P. Kasture and K. Shirsath, "Enhancing stock market prediction: a hybrid RNN-LSTM framework with sentiment analysis," *Indian Journal of Science and Technology,* vol. 17, pp. 1880-1888, 2024-01-01 2024.

[19] T. Dimpfl and S. Jank, "Can internet search queries help to predict stock market volatility?" *European financial management,* vol. 22, pp. 171-192, 2016-01-01 2016.

[20] R. P. Schumaker, Y. Zhang, C. Huang, and H. Chen, "Evaluating sentiment in financial news articles," *Decision Support Systems,* vol. 53, pp. 458-464, 2012-01-01 2012.

[21] S. Chang, L. Hwang, C. Li, and M. Yao, "News sentiment and its effect on price momentum and sentiment momentum," *Int. J. Trade, Econ. Financ,* vol. 8, pp. 251-257, 2017-01-01 2017.

User Satisfaction in AI-Driven Islamic Fintech: An Extended Technology Acceptance Model with Task–Technology Fit and Sharia Compliance

Mardiana Andarwati[1]*, Sari Yuniarti[2], Andriyan Rizki Jatmiko[3],
Firnanda Al-Islama Achyunda Putra[4], Galandaru Swalaganata[5], Ahmad Taufiq Andriono[6]
Department of Information Systems-Faculty of Information Technology,
Universitas Merdeka Malang, Malang City, East Java, Indonesia [1,3,4,5,6]
Department of Finance and Banking-Faculty of Economics and Business,
Universitas Merdeka Malang, Malang City, East Java, Indonesia [2]

***Abstract*—The rapid development of digital financial services has transformed financial intermediation through improved access, transparency, and efficiency. In the Indonesian context, Islamic financial technology (fintech) offers an alternative aligned with Sharia principles, particularly through e-ijarah contracts that provide MSMEs with productive asset access without interest-bearing debt. This study aims to empirically evaluate the determinants of user satisfaction in adopting AI-based e-ijarah applications by extending the Technology Acceptance Model (TAM) with Task–Technology Fit (TTF), Sharia Compliance (SC), Trust in AI, and Perceived Risk (PR). A survey of 75 food and beverage MSMEs in East Java was analyzed using Partial Least Squares Structural Equation Modeling (PLS-SEM). The findings indicate that Perceived Ease of Use (PEOU) strongly influences Perceived Usefulness (PU), which in turn significantly affects Behavioral Intention (BI), Actual Use (AU), and User Satisfaction (EUCS). Trust in AI and TTF also play significant roles in enhancing PU and BI. Interestingly, SC shows a significant but negative effect on PU, highlighting a contextual gap between digital automation and perceptions of religious compliance. PR negatively impacts both BI and AU, while Age does not moderate usage behavior. The study contributes conceptually by integrating TAM, TTF, and Sharia compliance in a single framework, and practically by offering insights for fintech developers and regulators to improve system usability, trust, and compliance clarity.**

Keywords*—Task–Technology fit; sharia compliance; technology acceptance model; user satisfaction; AI; Islamic fintech; MSMEs

I. Introduction

The development of digital financial services has brought significant transformation in the global financial ecosystem, especially through increased access, transparency, and efficiency of financial intermediation. One of the innovations that plays an important role is financial technology (fintech), whose adoption is greatly influenced by the level of suitability of technology with user needs or Task-Technology Fit (TTF). TTF serves as a conceptual framework that bridges the fit between technological features and user tasks, thereby influencing perceived usefulness and encouraging behavioral intent towards system use [1], [2], [3].

In the Indonesian context, sharia-based fintech is present as an alternative in accordance with Islamic principles, in line with the provisions of DSN-MUI regarding ijarah contracts [4], PSAK 107 [5], and Government Regulation No. 44 of 1997 related to limited access to capital for Micro, Small, and Medium Enterprises (MSMEs). MSMEs play a key role in national economic growth, but access to productive financing is still hampered by strict credit requirements, high interest rates, and complex bureaucracy [6], [7]. This condition has prompted some MSMEs to switch to informal financing such as loan sharks, which actually worsens business resilience.

As a solution, ijarah contracts provide access to productive assets without having to be bound by interest-bearing debt, so that the initial capital burden can be reduced and business profits are allocated to pay the rent of assets. Data from the East Java Provincial Cooperative Office (2024) shows that there are 9.86 million MSME units, with the APINDO survey (2024) reporting that 59% or around 5.85 million MSMEs still need better access to production and technology assets [8], [9]. However, despite the growing importance of Islamic fintech, the key problem lies in how AI-driven automation can align with users' perception of sharia compliance and technological trust. The lack of empirical understanding regarding this relationship has limited both adoption and satisfaction levels among MSMEs.

The development of digital technology—especially the integration of artificial intelligence (AI)—provides an opportunity to strengthen the implementation of ijarah in the form of *e-ijarah* AI enables automation of financing feasibility analysis through credit scoring, risk assessment, and real-time contract management. The study by Lasmiatun & Manteghi [10] confirms that AI in Islamic finance is able to increase efficiency while expanding financial inclusion. Similarly, Syed Zahiruddin prove that AI innovation in Indonesia and Malaysia can strengthen risk management and improve operational efficiency of sharia products, including digital ijarah [11].

Accordingly, this study seeks to answer the following questions: 1) What factors significantly influence user satisfaction and behavioral intention in adopting AI-driven e-ijarah systems? 2) How do Sharia Compliance and Trust in AI affect perceived usefulness and behavioral intention?

*Corresponding author.

In addition to technology, other factors also influence the adoption of sharia fintech. Perceived ease of use (PEOU) plays an important role in shaping the perception of technological usefulness, which then influences the behavioral intention to use the system. Similarly, trust in AI contributes to increasing users' confidence in the reliability of the system, while sharia compliance is the main determinant of acceptance among Muslim users. On the other hand, perceived risk can lower usage intent due to concerns about data and transaction security. Building on these considerations, this study investigates an AI-based e-Ijarah application that supports contract approval decisions by evaluating financial conditions, business types, locations, and target markets to ensure eligibility according to fund providers' standards.

This research framework also places behavioral intention as a mediator that bridges the perception of technology with actual use. Furthermore, actual usage will affect the level of user satisfaction, which is measured based on the end-user computing satisfaction (EUCS) dimension. The age factor is also considered as a moderator, considering that technology usage preferences can vary between age groups.

This research is focused on the creative economy subsector, especially food and beverage MSMEs, because it is proven to be resistant to economic shocks [12], [13]. The model offered is in the form of AI-based fintech integration in *e-ijarah*, which involves investors (capital owners), MSMEs (tenants), and coaches (business mentors) as business sustainability supervisors.

Although the Technology Acceptance Model (TAM) and Task–Technology Fit (TTF) have been widely applied in fintech adoption studies, prior research has not yet integrated Sharia Compliance (SC) and Trust in AI (TR) simultaneously within the MSME context in Indonesia. Existing studies tend to examine TAM and TTF separately [20], or focus only on Sharia compliance in Islamic banking [23], without considering the role of artificial intelligence as a trust-enhancing factor. This creates a research gap, particularly in understanding how religious compliance and technological trust interact to influence user perceptions and satisfaction in Islamic fintech applications.

The present study addresses this gap by developing an extended TAM model that incorporates TTF, SC, TR, and Perceived Risk. The findings reveal a novel result: Sharia Compliance significantly but negatively affects Perceived Usefulness. This contradicts much of the existing literature, which generally assumes that higher religious compliance should increase perceived usefulness and satisfaction [23]. The negative direction suggests a contextual perception gap, where MSME users may view automation and AI-driven processes as less transparent or less aligned with their expectations of Sharia principles. This theoretical contribution highlights the need for further refinement in designing AI-based Islamic fintech systems to ensure both usability and compliance clarity. Conceptually, this integration provides a more comprehensive framework for Islamic fintech adoption; practically, it offers guidance for regulators and developers in aligning system design with both technological and religious expectations.

Unlike previous fintech studies that merely extend TAM or TTF with general trust or risk constructs, this study introduces *Sharia Compliance* and *Trust in AI* as interdependent dimensions. Theoretically, this integration redefines behavioral pathways by linking ethical–religious assurance with algorithmic trust, offering a dual validation mechanism for user intention and satisfaction in Islamic fintech adoption.

Therefore, this study aims to empirically test an extended TAM model that incorporates TTF, SC, Trust in AI, and PR to explain user satisfaction with e-ijarah among MSMEs in East Java. Specifically, the study aims to test an extended Technology Acceptance Model (TAM) integrating TTF, SC, Trust in AI, and Perceived Risk to explain user intention, actual use, and satisfaction among MSMEs in East Java. This research is significant both theoretically—by introducing ethical and algorithmic dimensions into TAM—and practically—by guiding Islamic fintech developers to improve transparency and compliance clarity in e-ijarah systems.

The remainder of this paper is organized as follows: Section II reviews the literature study; Section III describes the research method; Section IV presents results and discussion; and Section V concludes with implications and future directions.

II. Literature Study

This study builds upon two foundational frameworks, Technology Acceptance Model (TAM) and Task-Technology Fit (TTF) to investigate the determinants of user intention and satisfaction in the context of e-Ijarah systems. TAM provides the theoretical basis for how Perceived Ease of Use (PEOU) and Perceived Usefulness (PU) shape behavioral intention and actual usage [14], while TTF emphasizes the alignment between user tasks and system capabilities [15]. In addition, this research incorporates context-specific constructs such as Trust in AI, Sharia Compliance, and Perceived Risk, which are particularly salient in Islamic fintech adoption. The outcome variable user satisfaction is assessed holistically, inspired by satisfaction models in information systems literature.

A. Technology Acceptance Model (TAM)

TAM posits that technology acceptance is primarily driven by two beliefs: PEOU and PU, which influence behavioral intention and actual use. Recent studies continue to validate TAM in fintech contexts. For example, [16] found that PEOU and PU significantly affect online purchase intention, with perceived risk having a negative effect. Furthermore, it has been demonstrated in an AI-enhanced e-shopping environment that PU had a stronger influence than PEOU, and trust emerged as a key predictor of user intention [17], [18]. These findings justify the inclusion of PEOU, PU, trust, and risk in the present model.

B. Task-Technology Fit (TTF)

TTF theory asserts that users are more likely to adopt technology when its functionality aligns well with their task requirements [19]. Recent empirical work among fintech and digital service professionals affirms that task-technology fit significantly predicts perceived usefulness and intention to use, it was found that combining TAM with TTF enhanced explanatory power [20], [21]. Accordingly, our model integrates TTF to capture the degree to which the e-Ijarah application supports users' specific tasks.

C. Context-Specific Variables: Trust, Sharia Compliance and Perceived Risk

Trust in AI is increasingly recognized as pivotal in acceptance of automated Islamic finance tools [17], [18]. Perceived Risk often predicts behavioral hesitance in online shopping adoption [22], [23]. Meanwhile, Sharia Compliance, although generally assumed positive, has yielded mixed results, with findings showing that it significantly influenced satisfaction but did not consistently strengthen the PU–trust relationship [24].

D. User Satisfaction (Holistic Perspective)

Instead of a full EUCS model, this study adopts a holistic satisfaction construct shaped by system performance, ease of use, and task alignment mirroring satisfaction frameworks frequently used in fintech and e-service research. The literature consistently underscores satisfaction as an important outcome following actual system use and perceived utility [25], [26].

E. Empirical Evidence from Recent Acceptance Studies

The empirical evidence summarized in Table I presents various recent studies that have applied acceptance models such as TAM, TTF, and external constructs in different contexts. The comparison highlights both consistency and divergence across findings. For example, variables such as Perceived Usefulness (PU) and Perceived Ease of Use (PEOU) have consistently shown significant effects on Behavioral Intention (BI) and Actual Use (AU) across most studies, confirming their strong theoretical foundation in TAM. On the other hand, several studies reported non-significant relationships for constructs such as Trust or Perceived Risk, which suggests that their influence may depend on contextual factors such as industry sector, cultural background, or technology maturity.

Although prior studies examined trust, risk, and sharia compliance in fintech adoption, limited research focuses on AI-driven decision-making in Islamic contracts. This study addresses that gap by analyzing acceptance of an AI-based e-Ijarah system for contract approval.

In the current study, these inconsistencies are particularly important. Some variables that were previously found to be non-significant, such as Task-Technology Fit (TTF) or Sharia Compliance (SC), show strong and significant effects in this research, indicating that different user groups (e.g., MSMEs in East Java) may evaluate technology differently compared to users in other contexts. Conversely, if certain variables that were significant in prior studies appear non-significant in this study, it emphasizes the importance of considering contextual variations in technology adoption. Therefore, the literature review not only strengthens the theoretical basis of this research but also highlights the novelty and contribution of the current findings by comparing them with prior evidence.

Building on the inconsistencies shown in prior studies (Table I), this study extends the TAM framework by adding TTF, SC, TR, and PR to capture the unique adoption dynamics of Islamic fintech.

TABLE I. COMPARATIVE ANALYSIS OF ACCEPTANCE MODELS OVER RECENT FIVE YEARS

Author(s) & Model Type	Included Constructs	R Square	Insignificant Variable
Imam Suroso et al. TAM + Perceived Risk [22]	PEOU, PU, Risk, Trust → Intention to Use Fintech (Indonesia)	$R^2 \approx 0.60$	Trust→ Intention not significant
Nagy & Hajdu. TAM + Trust [18]	PEOU, PU, Trust → Behavioral Intention (AI shopping, Hungary)	Not explicitly reported	Risk was excluded; Trust dominant predictor
Usman et al. TAM + Sharia Compliance [24]	PEOU, PU, Sharia Compliance, Trust → Satisfaction & Intention	R^2 Satisfaction ≈ 0.45	PU→ Satisfaction not always significant
Seyum et al. TTF + TAM [20]	TTF dimensions, PU, PEOU → Behavioral Intention (FinTech, Ethiopia)	R^2 not stated	Overfit TTF → PU not significant; Underfit TTF → PU negative
Alrawad et al. UTAUT2 + Trust Model [27]	Performance Exp., Effort Exp., Trust, Risk → Intention to Use FinTech	High R^2 (not specified)	Hedonic motivation not significant
Patnaik et al. Extended TAM (FinTech India) [28]	PEOU, PU, Trust, Satisfaction → Behavioral Intention	R^2 BI (not specified)	—
Taufiq-Hail et al. FinTech Satisfaction Model [29]	PEOU, Trust, Quality → Satisfaction → Loyalty (FinTech, Oman)	R^2 Loyalty (not specified)	Some trust dimensions not significant
Uddin & Nasrin FinTech Literacy Model [30]	FinTech Literacy, Satisfaction → Sustainable Intention	R^2 Sustainable Intention (not specified)	—
Taneja et al. Sustainability & FinTech [31]	FinTech Implementation, Satisfaction, Trust → Sustainability	R^2 Sustainability (not specified)	—
Perwitasari TAM (MSMEs FinTech Indonesia) [32]	PEOU, PU → Behavioral Intention to Use FinTech	R^2 BI (not specified)	—

III. RESEARCH METHODOLOGY

This research model shown in Fig. 1 proposes a conceptual model that integrates constructs from the Technology Acceptance Model (TAM), Task-Technology Fit (TTF), Sharia Compliance theory, Perceived Risk theory, and User Satisfaction to examine the acceptance and satisfaction of e-Ijarah technology. The model also introduces age as a moderating variable.

The model consists of five exogenous variables, two mediators, two endogenous variables, and one moderator, with the following structure:

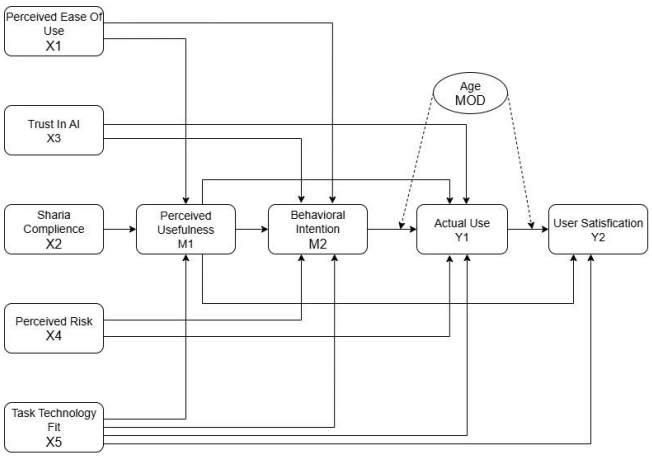


Fig. 1. Research model.

Exogenous Variables

X1: Perceived Ease of Use (PEOU)

X2: Sharia Compliance (SC)

X3: Trust in AI (TR)

X4: Perceived Risk (PR)

X5: Task-Technology Fit (TTF)

Mediating Variables

M1: Perceived Usefulness (PU) acts as a bridge between technological, risk, trust, and religious aspects toward user intention.

M2: Behavioral Intention (BI) mediates the relationship between perceptions and actual use behavior.

Endogenous Variables

Y1: Actual Use (AU), which reflects how often or consistently users interact with the e-Ijarah platform.

Y2: User Satisfaction (EUCS), which evaluates user experience across dimensions such as ease of use, accuracy, timeliness, format, and content.

Moderator Variable

Age moderates the influence of Behavioral Intention (BI) and Actual Use (AU) toward User Satisfaction. It is expected that user age may influence how intention translates into action, and how use translates into satisfaction.

H1: Perceived Ease of Use (PEOU) has a positive effect on Perceived Usefulness (PU).

H2: Sharia Compliance (SC) has a positive effect on Perceived Usefulness (PU).

H3: Trust in AI (TR) has a positive effect on Perceived Usefulness (PU).

H4: Perceived Risk (PR) has a negative effect on Perceived Usefulness (PU).

H5: Task-Technology Fit (TTF) has a positive effect on Perceived Usefulness (PU).

H6: Perceived Usefulness (PU) has a positive effect on Behavioral Intention (BI).

H7: Perceived Ease of Use (PEOU) has a positive effect on Behavioral Intention (BI).

H8: Sharia Compliance (SC) has a positive effect on Behavioral Intention (BI).

H9: Trust in AI (TR) has a positive effect on Behavioral Intention (BI).

H10: Perceived Risk (PR) has a negative effect on Behavioral Intention (BI).

H11: Task-Technology Fit (TTF) has a positive effect on Behavioral Intention (BI).

H12: Behavioral Intention (BI) has a positive effect on Actual Use (AU).

H13: Actual Use (AU) has a positive effect on User Satisfaction (EUCS).

H14: Perceived Usefulness (PU) has a positive effect on User Satisfaction (EUCS).

H15: Age moderates the relationship between Behavioral Intention (BI) and Actual Use (AU).

H16: Age moderates the relationship between Actual Use (AU) and User Satisfaction (EUCS).

The research population consisted of 93 food and beverage entrepreneurs classified as MSMEs in a district of East Java. Considering the total population size and the required confidence level of 95% with a 5% margin of error, the sample size was determined using the Slovin formula.

$$v = N/(1 + N) \quad (1)$$

$$v = 93/(1 + 93(0.05)2$$

$$v = 93/1.2325$$

$$v = 75$$

This calculation yielded a minimum sample of 75 respondents, which was deemed adequate and representative for statistical analysis using Partial Least Squares–Structural Equation Modeling (PLS-SEM), particularly since PLS-SEM is robust to relatively small sample sizes.

The inclusion criteria required that respondents: 1) were active owners or decision-makers of MSMEs, 2) operated within the food and beverage subsector of the creative economy, and 3) had experience using or considering digital financial applications, particularly AI-based e-Ijarah platforms. This ensured that the sample represented actual users or potential adopters of Islamic fintech systems, rather than general MSMEs without relevant exposure.

All 75 selected respondents completed the questionnaire fully, with no data excluded due to incompleteness. Data were analyzed using SmartPLS 4.0, which is widely employed in IS and fintech research due to its capability in handling complex models and relatively small sample sizes.

TABLE II. VARIABLES AND INDICATORS

Variable	Indicator	Code
Perceived Ease of Use (PEOU)	Easy to use	PEOU1
	Smooth interaction	PEOU2
	Easy to learn	PEOU3
	Quick to master	PEOU4
Perceived Usefulness (PU)	Helps tasks	PU1
	Increases productivity	PU2
	Improves performance	PU3
Trust in AI (TR)	Trust in AI system	TR1
	Privacy and security	TR2
	Comfortable using AI	TR3
Sharia Compliance (SC)	Follows Islamic principles	SC1
	Sharia-based process	SC2
	Complies with Sharia	SC3
Perceived Risk (PR)	Data misuse risk	PR1
	Security concern	PR2
	Online transaction worry	PR3
	Personal data risk	PR4
Task-Technology Fit (TTF)	Fits task requirements	TTF1
	Improves effectiveness	TTF2
	Aids task completion	TTF3
	Functional support	TTF4
Behavioral Intention (BI)	Intend to use regularly	BI1
	Recommend to others	BI2
	Continue using	BI3
Actual Use (AU)	Frequent use	AU1
	Used for various transactions	AU2
	Proper functionality usage	AU3
	Active utilization	AU4
User Satisfaction (EUCS)	Overall satisfaction	EUCS1
	Meets expectations	EUCS2
	Comfortable to use	EUCS3
	Satisfactory performance	EUCS4
	Enjoyable to use	EUCS5
Age (Moderator)	Respondent's age	Age

Table II presents the list of research variables and their associated indicators used in this study. Each variable was measured using multiple indicators adopted and adapted from relevant prior literature, aligned with the research model. These indicators were designed to reflect key constructs such as perceived ease of use, perceived usefulness, trust in AI, sharia compliance, perceived risk, task-technology fit, behavioral intention, actual system usage, and user satisfaction. The indicators were labeled using concise codes for easier referencing during the data analysis process. All items were measured using a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). All indicators were adapted from prior validated scales with minor contextual adjustments.

Perceived Risk (PR) is a negatively oriented construct, meaning higher scores reflect greater concern or fear regarding the system. Therefore, during data preprocessing, the scores for this variable were reverse-coded to maintain consistency in the interpretation of positive path relationships across the model. The moderator variable, age, was recorded as a continuous demographic input.

IV. RESULTS AND DISCUSSION

A. Result

1) Data demographic: This research involved 75 respondents who are owners or decision-makers of MSMEs in the food and beverage sector located in a district in East Java. The demographic profile of respondents is summarized based on several categories including gender, age, level of education, business duration, and role in business operations. Details are shown in Table III.

TABLE III. DEMOGRAPHIC RESPONDENTS

Category	Subcategory	Freq	(%)
Gender	Male	41	54.7%
	Female	34	45.3%
Age Group	< 25 years	8	10.7%
	26–35 years	21	28.0%
	36–45 years	27	36.0%
	> 45 years	19	25.3%
Education Level	Junior High or below	5	6.7%
	Senior High School	28	37.3%
	Diploma/ Bachelor's Degree	37	49.3%
	Postgraduate Degree	5	6.7%
Business Duration	< 1 year	6	8.0%
	1–3 years	22	29.3%
	4–6 years	28	37.3%
	> 6 years	19	25.3%
Role in Business	Owner	53	70.7%
	Manager/Operator	17	22.7%
	Others	5	6.6%

2) Measurement model analysis (outer model): Fig. 2 illustrates the measurement model and structural relationships among latent constructs. All outer loadings exceed the recommended threshold of 0.70, indicating that the indicators are strongly related to their latent constructs [33].

The path coefficients shown between latent variables represent the strength and direction of relationships. Most paths are positive and significant, although some show weak or negative relationships, such as the effect of Perceived Risk (PR) on Actual Use (AU), which shows a negative coefficient (-0.288), indicating that as users perceive higher risk, their

likelihood of using the system decreases highlighting a critical barrier that needs to be addressed in system adoption strategies.

Each latent construct also displays an R^2 value in the center, showing how much variance is explained by its predictors [33]. For instance, the construct User Satisfaction has an R^2 of 0.881, indicating that 88.1% of the variance in user satisfaction can be explained by the model.

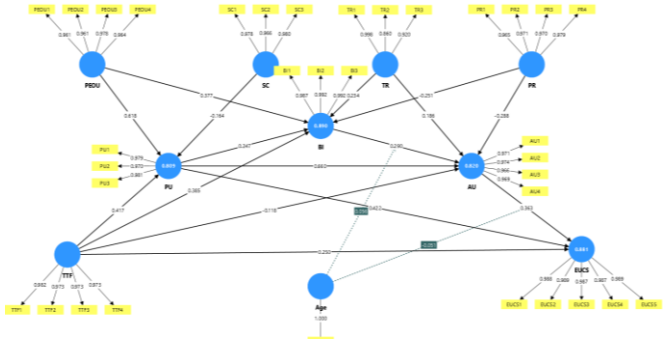


Fig. 2. Output line diagram.

Table IV shows convergent validity assesses whether the indicators truly reflect their intended constructs. This is evaluated through the Average Variance Extracted (AVE). An AVE value of 0.50 or higher is considered acceptable, indicating that the construct explains more than 50% of the variance in its indicators [34].

Our model explains acceptance of an AI-based e-Ijarah application that recommends contract approval by assessing financial condition, business type, location, and target market against funders' standards. In this context, TTF represents alignment between users' evaluation tasks and the AI decision logic; TAM constructs (PEOU, PU) capture usability and utility perceived from AI assistance. We further include Trust, Perceived Risk, and Sharia Compliance as context-specific determinants of intention and satisfaction, with the expectation that trust and sharia compliance enhance adoption, while perceived risk diminishes it.

Table V presents the reliability results of each construct using Cronbach's Alpha and Composite Reliability. These two indicators are used to assess internal consistency of the measurement model.

All constructs exhibit Cronbach's Alpha values above 0.70, indicating that the items within each construct are consistently measuring the same underlying concept. Similarly, the Composite Reliability values also exceed the recommended threshold of 0.70, confirming that each construct demonstrates strong internal consistency and measurement accuracy.

The highest reliability is observed in User Satisfaction (EUCS) with a Cronbach's Alpha of 0.994 and rho_c of 0.995, while the lowest, yet still acceptable, reliability is seen in Trust in AI (TR) with a Cronbach's Alpha of 0.963 and rho_c of 0.949.

TABLE IV. CONVERGENT VALIDITY OF CONSTRUCTS

Indicator	Loading Factor	Valid / No	AVE
PEOU1	0.961	Valid	0.934
PEOU2	0.961	Valid	
PEOU3	0.978	Valid	
PEOU4	0.964	Valid	
PU1	0.979	Valid	0.954
PU2	0.970	Valid	
PU3	0.981	Valid	
SC1	0.978	Valid	0.950
SC2	0.966	Valid	
SC3	0.980	Valid	
TR1	0.998	Valid	0.860
TR2	0.860	Valid	
TR3	0.920	Valid	
PR1	0.965	Valid	0.944
PR2	0.971	Valid	
PR3	0.970	Valid	
PR4	0.979	Valid	
BI1	0.987	Valid	0.980
BI2	0.992	Valid	
BI3	0.992	Valid	
AU1	0.971	Valid	0.941
AU2	0.974	Valid	
AU3	0.966	Valid	
AU4	0.969	Valid	
TTF1	0.982	Valid	0.951
TTF2	0.973	Valid	
TTF3	0.973	Valid	
TTF4	0.973	Valid	
EUCS1	0.988	Valid	0.976
EUCS2	0.989	Valid	
EUCS3	0.987	Valid	
EUCS4	0.987	Valid	
EUCS5	0.989	Valid	

*All constructs exceed the AVE threshold, confirming that convergent validity is achieved across all latent variables.

TABLE V. RELIABILITY ANALYSIS OF CONSTRUCTS

Variable	Cronbach's Alpha	Composite Reliability (ρc)
PEOU	0.976	0.984
PU	0.976	0.984
SC	0.974	0.984
TR	0.963	0.949
PR	0.980	0.985
BI	0.990	0.994
AU	0.979	0.986
TTF	0.983	0.990
EUCS	0.994	0.995

3) Structural model analysis (inner model): The inner model or structural model analysis aims to evaluate the relationships between latent variables based on the proposed hypotheses. The evaluation includes the analysis of path coefficients, R-square values, F-square effect sizes, and discriminant validity using the Fornell-Larcker criterion.

R-square (R^2) indicates how much of the variance in the dependent variable is explained by the independent variables. All endogenous variables have $R^2 > 0.75$, indicating the model has substantial explanatory power for Actual Use, Behavioral Intention, Perceived Usefulness, and User Satisfaction.

Table VI shows the R^2 values are all above the 0.75 threshold, which is considered substantial according to Hair et al. (2017). Specifically, Behavioral Intention (0.890), Actual Use (0.820), Perceived Usefulness (0.809), and User Satisfaction (0.881) are strongly predicted by their respective independent variables. These results confirm that the proposed model has high explanatory power and is effective in capturing the key factors influencing e-Ijarah application usage and satisfaction.

TABLE VI. R-SQUARE VALUE

Variable	R-Square	Result
AU	0.820	Strong
BI	0.890	Strong
PU	0.809	Strong
User Satisfaction	0.881	Strong

The detailed path coefficients and their statistical significance are summarized in Table VII, which presents the key relationships and hypothesis testing results. The F-square (f^2) values measure the effect size or contribution of each exogenous variable to the R^2 value of an endogenous variable when included in the model [35]. Values of 0.02, 0.15, and 0.35 represent small, medium, and large effects, respectively [36].

TABLE VII. F-SQUARE VALUE

Path	F^2 Value	Result
PEOU → PU	0.972	Strong
PR → BI	0.544	Strong
PU → AU	0.558	Strong
TTF → BI	0.632	Strong
PU → User Satisfaction	0.353	Strong
PEOU → BI	0.322	Medium
PU → BI	0.112	Weak
TTF → PU	0.293	Medium
AU → User Satisfaction	0.392	Strong
TTF → User Satisfaction	0.239	Medium
PR → AU	0.316	Medium
TR → BI	0.486	Strong
TR → AU	0.134	Weak
SC → PU	0.076	Weak
TTF → AU	0.022	Weak
BI → AU	0.066	Weak
PU → BI	0.112	Weak
Age interactions	< 0.03	Negligible

The path analysis presents the estimated relationships between latent variables in the structural model, indicated by path coefficients (β), along with their statistical significance. A positive and statistically significant coefficient (T-statistic $\geqslant$ 1.96 and p-value $\leqslant$ 0.05) supports the proposed hypothesis, indicating a meaningful influence of one construct on another. Conversely, non-significant paths suggest that the hypothesized relationships are not empirically supported in the model [37].

TABLE VIII. PATH COEFFICIENT RESULTS

Path	T-Stat	P-Val	Result
AU→User Satisfaction	4.609	0.000	Significant
BI → AU	2.152	0.031	Significant
PEOU → BI	4.583	0.000	Significant
PEOU → PU	9.610	0.000	Significant
PR → AU	4.495	0.000	Significant (negative)
PR → BI	3.972	0.000	Significant (negative)
PU → AU	5.811	0.000	Significant
PU → BI	2.720	0.007	Significant
PU → User Satisfaction	4.950	0.000	Significant
SC → PU	2.742	0.006	Significant (negative)
TR → AU	2.640	0.008	Significant
TR → BI	3.735	0.000	Significant
TTF → AU	1.175	0.240	Not Significant
TTF → BI	6.288	0.000	Significant
TTF → PU	4.702	0.000	Significant
TTF→ User Satisfaction	4.097	0.000	Significant
Age → AU	0.698	0.486	Not Significant
Age →User Satisfaction	0.305	0.760	Not Significant
Age × AU →User Satisfaction	1.146	0.252	Not Significant
Age × BI → AU	1.145	0.252	Not Significant

Table VIII presents several key relationships, such as those between Perceived Ease of Use and Perceived Usefulness, Task-Technology Fit and Behavioral Intention, as well as Perceived Risk and Actual Use are supported, revealing strong theoretical alignment and practical relevance in the context of e-Ijarah adoption.

B. Discussion

The study reveals a strong and significant positive relationship between PEOU and PU ($T = 9.610$, $p < 0.001$, $f^2 = 0.972$), consistent with the core assumptions of the Technology Acceptance Model (TAM) by Davis (1989). When users perceive an application as easy to use, they are more likely to perceive it as useful. This aligns with previous studies [38], reinforcing that usability is a fundamental driver of utility perception in digital financial tools.

PU significantly influences both Behavioral Intention (BI) ($T = 2.720$, $f^2 = 0.112$) and Actual Use (AU) ($T = 5.811$, $f^2 = 0.558$), supporting its central role in TAM. PU also significantly impacts User Satisfaction ($T = 4.950$, $f^2 = 0.353$). These findings confirm that users' perception of utility affects not only their intention to use but also their experience and satisfaction after actual usage. This result is in line with research [39].

BI positively affects AU ($T = 2.152$, $f^2 = 0.066$), showing that users with stronger intentions tend to actually use the system. However, the effect size is small, indicating that other factors such as PU and PR play a more dominant role in predicting actual usage in the e-Ijarah context.

TTF is a strong predictor of BI ($T = 6.288$, $f^2 = 0.632$) and PU ($T = 4.702$, $f^2 = 0.293$), supporting the model proposed by Goodhue & Thompson (1995). TTF also significantly affects User Satisfaction ($T = 4.097$, $f^2 = 0.239$), showing that alignment between task requirements and system capabilities enhances the user experience. However, its influence on AU is not significant ($T = 1.175$), suggesting that even with good task-system fit, external factors may inhibit actual usage behavior.

Trust (TR) significantly influences both BI ($T = 3.735$, $f^2 = 0.486$) and AU ($T = 2.640$, $f^2 = 0.134$). This finding highlights that trust in AI-based e-Ijarah systems is essential for encouraging adoption, especially in contexts involving sensitive data and financial transactions.

In contrast, Perceived Risk (PR) has a negative and significant effect on both BI ($T = 3.972$, $f^2 = 0.544$) and AU ($T = 4.495$, $f^2 = 0.316$). This supports prior findings by [23][40], emphasizing that minimizing perceived risk is critical in digital Islamic finance.

Actual Use (AU) significantly improves User Satisfaction ($T = 4.609$, $f^2 = 0.392$), indicating that continued use and engagement with the e-Ijarah platform positively impact user evaluation. This confirms findings from the EUCS framework and prior studies such as.

Interestingly, Sharia Compliance (SC) shows a significant but negative impact on PU ($T = 2.742$, $f^2 = 0.076$). This unexpected result may indicate a perception gap between automation (AI features) and sharia principles. Users may feel that digitalization compromises transparency or compliance clarity, requiring a re-evaluation of how sharia features are communicated and implemented in the application.

Theoretically, the extended model contributes by linking Sharia Compliance and Trust in AI as dual antecedents of perceived usefulness, offering a novel ethical–technological integration. This re-specifies TAM pathways by demonstrating that compliance clarity and algorithmic transparency jointly predict adoption in Islamic fintech—an advancement over prior single-dimensional trust frameworks.

The negative effect of SC on PU may reflect skepticism toward the transparency of AI-driven processes. Users might perceive that digital automation reduces the clarity of Sharia compliance, which calls for more explicit communication of compliance features in system design.

The moderating effects of Age and its interactions (e.g., Age × BI → AU, Age × AU → User Satisfaction) are not statistically significant (all $T < 1.96$, $f^2 < 0.03$). This indicates that age does not substantially alter user behavior in this context, supporting the idea that digital behavior is becoming more uniform across age groups in modern MSMEs.

From a theoretical standpoint, this study contributes to the integration of TAM, TTF, EUCS, and Islamic compliance in a single model, enriching IS and fintech literature. Practically, the results suggest that improving usability, aligning system design with tasks, and building user trust are key to adoption. Moreover, developers of Islamic fintech applications must address users' perception of compliance clarity to ensure trust and utility perceptions are aligned.

As an internal validation, the current results were compared with recent fintech adoption models by Usman et al., [41] and Seyum et al., [20]. The comparable R^2 (>0.80) and consistent significant paths for PU–BI and BI–AU confirm structural robustness. Nonetheless, future studies could empirically validate this model using cross-sample or experimental design to strengthen external validity.

1) Practical implications: The findings highlight key implications for developers and policymakers in Islamic fintech. The negative effect of Sharia Compliance (SC) on Perceived Usefulness (PU) reflects a perception gap between automation and religious assurance. Developers should include clear Sharia indicators—such as certification icons, transparent transaction logs, and short explanations of ijarah procedures—to strengthen users' confidence. Enhancing AI transparency through explainable dashboards and improving users' digital literacy via training or awareness programs are also essential. Furthermore, unified national standards for AI-based Sharia compliance are recommended to ensure consistency, reduce interpretational gaps, and promote sustainable trust in AI-driven e-ijarah systems. Regulators should standardize compliance communication to minimize user uncertainty regarding automation in religious contracts.

V. Conclusion, Limitations and Future Research Directions

This study investigated the determinants of user satisfaction in adopting AI-based applications, incorporating variables such as Perceived Ease of Use, Perceived Usefulness, Behavioral Intention, Actual Use, Trust in AI, Sharia Compliance, Perceived Risk, and Task-Technology Fit, with Age serving as a moderating variable. The analysis utilized Partial Least Squares Structural Equation Modeling (PLS-SEM) and involved a sample of 75 respondents from MSMEs in East Java.

The results showed that most of the proposed hypotheses were supported. Perceived Usefulness emerged as a key factor influencing both Behavioral Intention and Actual Use, while Actual Use significantly affected User Satisfaction. Trust in AI and Task-Technology Fit were also found to be significant contributors to perceived usefulness and satisfaction. Interestingly, Sharia Compliance had a negative but significant influence on Perceived Usefulness, suggesting a potential

conflict or misunderstanding regarding religious compliance and perceived utility. Age, as a moderating variable, did not have a significant effect on either Actual Use or User Satisfaction.

The model demonstrated strong explanatory power with high R-square values (e.g., 0.881 for User Satisfaction and 0.820 for Actual Use), and most constructs showed strong reliability and convergent validity. Effect size (f^2) analysis further confirmed that several paths had medium to large effects, particularly from Perceived Usefulness and Actual Use toward satisfaction.

In conclusion, this study highlights the critical roles of perceived benefits, ease of use, trust, and technological alignment in shaping users' behavioral intentions and satisfaction toward AI-based applications in MSMEs. These insights can guide future application development and policy formulation to foster broader acceptance, especially within culturally and religiously sensitive user groups.

This study is limited by its small sample size (n = 75) and single-sector focus (food and beverage MSMEs in East Java), which may restrict generalizability. Future research should include cross-sector and cross-regional samples, employ mixed-method validation such as multi-group analysis or experimental AI-trust scenarios, and compare e-ijarah with other Islamic contracts (e.g., murabahah, mudharabah) to enhance external validity.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial support from the Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia (Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi) under the 2025 Research Grant Scheme (SP DIPA 139.04.1.693320/2025). This support made the completion of this study possible.

REFERENCES

[1] P. Wijayanti, I. S. Mohamed, and D. Daud, "Computerized accounting information systems: An application of task technology fit model for microfinance," International Journal of Information Management Data Insights, vol. 4, no. 1, p. 100224, 2024.

[2] F. Makoza, "Leveraging Distributed ledger technology in Unified beneficiary registry system of Malawi: A Task-technology fit Theory perspective," Working Paper ICTD, 2025.

[3] L. Zavolokina, M. Dolata, and G. Schwabe, "FinTech transformation: How IT-enabled innovations shape the financial sector," in International Workshop on Enterprise Applications and Services in the Finance Industry, Springer, 2016, pp. 75–88.

[4] Dewan Syari'ah Nasional, Fatwa No: 09/DSN-MUI/IV/2000 Tentang Pembiayaan IJARAH. Indonesia: MUI, 2020.

[5] IAI, DRAF EKSPOSUR PSAK 107 : Akuntansi Ijarah (Revisi 2020). Jakarta: Graha Akuntan, 2020.

[6] M. Andarwati and G. Swalaganata, "Analysis of Promotional Media Selection Based on Modified Analytical Hierarchy Process to Increase Halal Products Sales Volume," IQTISHODUNA: Jurnal Ekonomi Islam, vol. 12, no. 1, pp. 1–20, Apr. 2023, doi: 10.54471/iqtishoduna.v12i1.2270.

[7] M. Andarwati, G. Swalaganata, and N. D. Hendrawan, "Implementation of E-Mudharabah Application as A Learning Media for Islamic Economics Accounting Using Formative Assessments Method," in 2023 9th International Conference on Education and Technology (ICET), IEEE, 2023, pp. 116–122.

[8] S. A. Suryaningsih, "Aplikasi Mudharabah dalam Perbankan Syariah di Indonesia," Journal of Innovation in Business and Economics, vol. 4, no. 1, pp. 13–24, 2013.

[9] R. A. Ghofur, "Konstruksi akad dalam pengembangan produk perbankan syariah di indonesia," Al-'Adalah, vol. 12, no. 1, pp. 493–506, 2015.

[10] K. M. T. Lasmiatun and N. Manteghi, "The Impact of Artificial Intelligence (AI) Implementation on Islamic Financial Literacy and Global Economic Changes in the Banking World," Journal of Islamic Economics and Bussines Ethics, vol. 2, no. 1, pp. 23–43, 2025.

[11] S. Z. bin Syed Musa, "Harnessing Artificial Intelligence for Optimized Zakat and Waqf Management: Strategic Insights from Indonesia and Malaysia," INTERNATIONAL JOURNAL OF CONTEMPORARY ISSUES (IJCI), p. 237.

[12] M. Andarwati et al., "Adoption of Generative AI-Enhanced Profit Sharing Digital Systems in MSMEs: A Comprehensive Model Analysis.," International Journal of Advanced Computer Science & Applications, vol. 16, no. 1, 2025.

[13] M. Andarwati, G. Swalaganata, F. Y. Pamuji, and N. D. Hendrawan, "Development of an e-Mudharabah Website Using the System Usability Scale (SUS) Method," KnE Social Sciences, vol. 9, no. 29, pp. 43–55, 2024.

[14] A. Raza and T. Tursoy, "Technology Acceptance Model and Fintech: An Evidence from Italian Banking Industry," Revista Mexicana de Economia y Finanzas Nueva Epoca, vol. 20, no. 1, Jan. 2025, doi: 10.21919/remef.v20i1.993.

[15] P. Wijayanti, I. S. Mohamed, and D. Daud, "Computerized accounting information systems: An application of task technology fit model for microfinance," International Journal of Information Management Data Insights, vol. 4, no. 1, p. 100224, 2024, doi: https://doi.org/10.1016/j.jjimei.2024.100224.

[16] M. A. Jamshaid, A. Rashid, and F. Idrees, "Digital Banking Adoption in Emerging Economies: Revisiting TAM through the Lens of Perceived Risk," Qlantic Journal of Social Sciences, vol. 6, no. 2, pp. 206–216, Jun. 2025, doi: 10.55737/qjss.vi-ii.25366.

[17] W. Zhang, S. Siyal, S. Riaz, R. Ahmad, M. F. Hilmi, and Z. Li, "Data Security, Customer Trust and Intention for Adoption of Fintech Services: An Empirical Analysis From Commercial Bank Users in Pakistan," Sage Open, vol. 13, no. 3, Jul. 2023, doi: 10.1177/21582440231181388.

[18] S. Nagy and N. Hajdú, "Consumer Acceptance of the Use of Artificial Intelligence in Online Shopping: Evidence From Hungary," Amfiteatru Economic, vol. 23, no. 56, pp. 1–1, 2021, doi: 10.24818/EA/2021/56/155.

[19] J. Mantik et al., "Does task technology fit, social influence, and habit impact on actual usage of mobile banking in Indonesia?," Online, 2023.

[20] A. A. Seyum, S. Wang, N. Zhang, and L. Wang, "Fit for the future: Examining the impact of task-technology fit on bank employee intentions to use FinTech," vol. 28, no. 1, 2025.

[21] B. Dash, P. Sharma, and S. Swayamsiddha, "Organizational Digital Transformations and the Importance of Assessing Theoretical Frameworks such as TAM, TTF, and UTAUT: A Review," 2023. [Online]. Available: www.ijacsa.thesai.org

[22] I. Suroso, M. F. Afandi, and A. Galushasti, "Does perceived risk? A study of technology acceptance model on online shopping intention," Academy of Strategic Management Journal, vol. 21, no. 3, pp. 1–12, 2022.

[23] M. Ali, S. A. Raza, B. Khamis, C. H. Puah, and H. Amin, "How perceived risk, benefit and trust determine user Fintech adoption: a new dimension for Islamic finance," Foresight, vol. 23, no. 4, pp. 403–420, 2021, doi: https://doi.org/10.1108/FS-09-2020-0095.

[24] H. Usman, N. W. K. Projo, C. Chairy, and M. G. Haque, "The exploration role of Sharia compliance in technology acceptance model for e-banking (case: Islamic bank in Indonesia)," Journal of Islamic Marketing, vol. 13, no. 5, pp. 1089–1110, 2022, doi: 10.1108/JIMA-08-2020-0230.

[25] N. Abdulaziz Alshathry and S. Abdullah Almeshal, "Fintech Perceived Usefulness, Ease Of Use Among Consumers And Its Effect On Satisfaction And Continuous Usage: An Impractical Study On STC Pay And Apple Pay In Saudi Arabia's Retail Sector," International Journal of Management & Information Technology, vol. 17, 2022, doi: 10.24297/ijmit.v17.

[26] D. Novira, H. S. Utomo, and I. H. Mulyanto, "Influence of Perceived Ease of Use and Perceived Usefulness towards Continuance Intention with Customer Satisfaction as Intervening Variable: a study of Startup Companies Using e-Wallet," Journal of Business Management and Economic Development, vol. 2, no. 02, pp. 602–614, Mar. 2024, doi: 10.59653/jbmed.v2i02.669.

[27] M. B. Amnas, M. Selvam, M. Raja, S. Santhoshkumar, and S. Parayitam, "Understanding the Determinants of FinTech Adoption: Integrating UTAUT2 with Trust Theoretic Model," Journal of Risk and Financial Management, vol. 16, no. 12, 2023, doi: 10.3390/jrfm16120505.

[28] A. Patnaik, P. Kudal, S. Dawar, V. Inamdar, and P. Dawar, "Exploring User Acceptance of Digital Payments in India: An Empirical Study Using an Extended Technology Acceptance Model in the Fintech Landscape," International Journal of Sustainable Development and Planning, vol. 18, no. 8, pp. 2587–2597, 2023, doi: 10.18280/ijsdp.180831.

[29] A. M. Taufiq-Hail et al., "Investigating the impact of customer satisfaction, trust, and quality of services on the acceptance of delivery services companies and related applications in Omani context: A Predictive model assessment using PLSpredict," Cogent Business and Management, vol. 10, no. 2, 2023, doi: 10.1080/23311975.2023.2224173.

[30] Md. K. Uddin and S. Nasrin, "The Mediating Effect of Customer Satisfaction on Fintech Literacy and Sustainable Intention of Using Mobile Financial Services," Open Journal of Business and Management, vol. 11, no. 05, pp. 2488–2504, 2023, doi: 10.4236/ojbm.2023.115138.

[31] S. Taneja, A. Siraj, L. Ali, A. Kumar, S. Luthra, and Y. Zhu, "Is FinTech Implementation a Strategic Step for Sustainability in Today's Changing Landscape? An Empirical Investigation," IEEE Trans Eng Manag, vol. 71, pp. 7553–7565, 2024, doi: 10.1109/TEM.2023.3262742.

[32] A. W. Perwitasari, "The Effect of Perceived Usefulness and Perceived Easiness towards Behavioral Intention to Use Fintech by Indonesian MSMEs," The Winners, vol. 23, no. 1, pp. 1–9, 2022, doi: 10.21512/tw.v23i1.7078.

[33] G. A. Putri, A. K. Widagdo, and D. Setiawan, "Analysis of financial technology acceptance of peer to peer lending (P2P lending) using extended technology acceptance model (TAM)," Journal of Open Innovation: Technology, Market, and Complexity, vol. 9, no. 1, p. 100027, 2023, doi: https://doi.org/10.1016/j.joitmc.2023.100027.

[34] J. D. Saville and L. L. Foster, "Does technology self-efficacy influence the effect of training presentation mode on training self-efficacy?," Computers in Human Behavior Reports, vol. 4, p. 100124, 2021, doi: https://doi.org/10.1016/j.chbr.2021.100124.

[35] J. Hair, M. Sarstedt, C. Ringle, and S. Gudergan, Advanced Issues in Partial Least Squares Structural Equation Modeling. 2017.

[36] J. Hair, G. T. M. Hult, C. Ringle, and M. Sarstedt, A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM). 2022.

[37] G. C. Altes, A. K. S. Ong, and J. D. German, "Determining factors affecting Filipino consumers' behavioral intention to use cloud storage services: An extended technology acceptance model integrating valence framework," Heliyon, vol. 10, no. 4, p. e26447, 2024, doi: https://doi.org/10.1016/j.heliyon.2024.e26447.

[38] C.-H. Huang, "Using PLS-SEM Model to Explore the Influencing Factors of Learning Satisfaction in Blended Learning," Educ Sci (Basel), vol. 11, no. 5, 2021, doi: 10.3390/educsci11050249.

[39] T. Zobeidi, S. B. Homayoon, M. Yazdanpanah, N. Komendantova, and L. A. Warner, "Employing the TAM in predicting the use of online learning during and beyond the COVID-19 pandemic," Front Psychol, vol. 14, 2023, doi: 10.3389/fpsyg.2023.1104653.

[40] K. Siregar, A. Soma, and P. Sitorus, "Factor affecting islamic fintech adoption by Gen Z in West Java with religion as moderating variable," Keynesia : International Journal of Economy and Business, vol. 3, pp. 129–140, Aug. 2024, doi: 10.55904/keynesia.v3i2.1277.

[41] H. Usman, N. W. K. Projo, C. Chairy, and M. G. Haque, "The exploration role of Sharia compliance in technology acceptance model for e-banking (case: Islamic bank in Indonesia)," Journal of Islamic Marketing, vol. 13, no. 5, pp. 1089–1110, 2022.

Multimodal Deep Learning for Tuberculosis Detection Using Cough Audio and Clinical Data with Health Acoustic Representations (HeAR)

Rinaldi Anwar Buyung*, Widi Nugroho
Department of Data Science, Seleris Meditekno Internasional, Jakarta, Indonesia

***Abstract*—Tuberculosis (TB) remains a significant global health challenge, necessitating rapid and accessible screening methods. This study proposes a multimodal deep learning model for non-invasive TB detection by fusing acoustic features from cough sounds with clinical metadata. We utilize the pre-trained Health Acoustic Representations (HeAR) model as a powerful backbone to extract features from mel-spectrograms of cough audio. These acoustic features are combined with clinical data, including sex, age, and key symptoms through a late-fusion architecture. The model was trained and evaluated on a balanced dataset of 16,000 samples derived from the CODA TB DREAM Challenge dataset. Our proposed multimodal approach achieved a high overall accuracy of 90% on the unseen test set, with balanced precision, recall, specificity, and F1-scores of 0.90 for both TB-positive and non-TB classes. These results demonstrate the effectiveness of using cough sound as a non-invasive vocal biomarker, amplified by combining advanced acoustic representations with clinical context. This highlights the potential of our method as a robust, low-cost, and scalable tool for early TB screening.**

***Keywords*—Tuberculosis; cough detection; Health Acoustic Representation; multimodal; vocal biomarker**

I. Introduction

Tuberculosis (TB) continues to be a devastating global health crisis. It is caused by the bacterium Mycobacterium tuberculosis, the disease primarily affects the lungs (pulmonary TB) and spreads through the air when an infected person coughs, sneezes, or speaks [1]. In 2023 alone, an estimated 10.8 million people fell ill with the disease, and it claimed the lives of 1.25 million people, reaffirming its status as a leading cause of death from a single infectious agent worldwide. The epidemic's immense scale underscores the critical need for global efforts to combat TB. These efforts align with the United Nations Sustainable Development Goal of ending the epidemic by 2030 [2]. The burden of this disease is unevenly distributed, with the majority of cases occurring in developing countries. Indonesia ranks as the country with the second-highest TB burden in the world after India, which underscores the urgency to develop more effective, rapid, and accessible screening and diagnostic methods within the country [3].

Conventional diagnostic methods for TB, such as sputum smear microscopy, bacterial culture, and rapid molecular tests (e.g., Xpert MTB/RIF), often have limitations. Sputum smear microscopy has variable sensitivity and may fail to detect cases with a low bacterial load. Bacterial culture, while being the gold standard, requires several weeks to yield results, which can delay treatment and increase the risk of transmission [4]. Meanwhile, molecular tests and radiological examinations like chest X-rays, although more accurate, require expensive laboratory infrastructure, specialized equipment, and expert personnel that are not always available in primary healthcare facilities or remote areas. These limitations hinder efforts for early detection and massive-scale containment of TB spread [5].

The pathological progression of pulmonary TB significantly alters the respiratory system's mechanics and acoustics. The formation of granulomas, inflammation of the airways, and accumulation of fluid or sputum change the physical properties of the lungs and trachea. These changes directly impact the sound produced during a cough, which is a forceful expulsion of air [6]. Consequently, coughs from TB-infected individuals can exhibit distinct acoustic patterns, such as variations in spectral energy, frequency components, and temporal characteristics compared to those from healthy individuals [7]. This acoustic signature provides a physiological basis for using cough sound as a non-invasive biomarker for TB screening [8].

Along with advancements in digital technology, sound analysis using artificial intelligence (AI) has emerged as a promising alternative solution. Deep learning, a subset of machine learning, enables systems to learn from data and understand complex concepts without explicit programming [9]. By learning from experience, computers can autonomously perform tasks, making them highly suitable for analyzing complex patterns in medical data. In previous studies, for instance, deep learning particularly Convolutional Neural Networks (CNNs) was successfully employed to detect diabetic retinopathy using retinal fundus images [10], [11]. Similarly, cough sounds, as a primary symptom of TB, contain rich acoustic information and can serve as a non-invasive vocal biomarker [12]. This capability offers the potential for a low-cost, rapid, and remotely deployable screening tool using common devices such as smartphones [13].

To maximize the potential of audio analysis, this study proposes the use of Health Acoustic Representations (HeAR), a foundation model developed by Google AI specifically for health-related sounds. HeAR is built on a transformer-based architecture and employs a self-supervised learning strategy, having been pre-trained on over 300 million two-second audio clips encompassing sounds like coughs and breaths. This

*Corresponding author.

extensive pre-training enables HeAR to generate powerful and generalized acoustic embeddings. In benchmark evaluations, simple linear classifiers trained on HeAR embeddings achieved state-of-the-art or competitive performance across various health tasks, including the detection of conditions like COVID-19 and the inference of smoking status from cough recordings [14].

By leveraging these rich representations as a backbone, this study aims to build a robust classifier for TB [15]. Furthermore, this study adopts a multimodal approach by not only relying on cough sounds but also integrating essential clinical data such as sex, age, and key symptoms (hemoptysis, night sweats, weight loss, and fever). The combination of advanced acoustic features from HeAR and contextual clinical information is expected to significantly improve detection accuracy and reliability, leading to a more holistic and accurate screening model [16].

This study makes several key contributions. First, we introduce a multimodal deep learning architecture that effectively fuses powerful pre-trained acoustic features from the HeAR model with essential clinical metadata. Second, we validate our approach on a large, balanced dataset, demonstrating its robustness and high accuracy. The remainder of this paper is organized as follows: Section II reviews related work in acoustic-based TB detection. Section III details the dataset, preprocessing steps, and our proposed model architecture. Section IV presents the experimental results and analysis. Finally, Section V discusses the clinical implications of our findings. Section VI concludes the study, and outlines directions for future research.

II. Related Work

The use of cough sound analysis as a low-cost, non-invasive tool for tuberculosis (TB) screening has been an active area of research. Early studies established the foundational potential of this approach by applying traditional machine learning models to handcrafted acoustic features. For instance, Botha et al. employed statistical classifiers like Logistic Regression on short-term spectral features combined with clinical data, achieving a sensitivity of 96% at a specificity of 72% on a dataset of 518 coughs. While pioneering, this work relied on manually engineered features (e.g., MFCC), which may not fully capture the complex, subtle acoustic patterns indicative of TB [17].

Subsequent research shifted towards deep learning to automate feature extraction and improve performance. Xu et al. developed a sophisticated fusion model using a Bidirectional Long Short-Term Memory (Bi-LSTM) network combined with a Convolutional Neural Network (CNN). Their approach, which fused traditional features like MFCC and Zero-Crossing Rate (ZCR) with features learned from spectrograms, reported a very high accuracy of 96.33%. However, this impressive result was obtained from a relatively small and homogenous dataset of 456 coughs from a limited number of participants. This raises important questions about the model's scalability and ability to generalize to more diverse, real-world populations [18].

In contrast, recent studies have demonstrated the power of simpler architecture when applied to large-scale datasets. Yadav et al. utilized a massive dataset of over 500,000 cough recordings and found that a simple 1D CNN trained on Mel-Frequency Cepstral Coefficients (MFCCs) achieved a high accuracy of 91%, exceeding the WHO's requirements for a screening test. This work underscores that with sufficient data, less complex models can be highly effective. Similarly, Kafentzis et al. also worked with a large, real-world dataset, achieving a respectable Area Under the Curve (AUC) of approximately 0.80 by combining audio features with clinical metadata, further highlighting the importance of a multimodal approach [19].

Addressing the need for larger and more diverse datasets, Kafentzis et al. utilized a substantial dataset of over 9,000 coughs collected "in the wild" via a mobile application. Using a combination of spectral features and clinical metadata, their statistical models achieved a respectable Area Under the Curve (AUC) of approximately 0.80 [20]. Their work highlighted the feasibility of large-scale data collection but also demonstrated that performance on noisy, real-world data remains a significant challenge, suggesting that more powerful feature representation is needed. A summary of these key studies is presented in Table I.

TABLE I. State-of-the-Art

Study	Methodology	Audio Features	Dataset	Evaluation Metrics
[17]	Logistic Regression	MFCC, Log Spectral Energies, Clinical Data	518 Coughs	Sensitivity = 96% Specificity = 72%
[18]	BiLSTM + Conv2D	MFCC, ZCR, RMS, Chroma Cens, Short Time Energy, Spectrogram	456 Coughs	Accuracy = 96.33% Specificity = 94.99%
[19]	1D CNN	MFCC	502,252 Coughs	Accuracy = 91% AUC = 0.8
[20]	CNN	LLDs, Clinical Data	9772 Coughs	AUC = 0.8

While these studies have established the potential of cough sound analysis for TB screening, clear research gaps remain. Previous works have either relied on handcrafted features or custom-trained CNNs. Furthermore, these investigations were often conducted on considerably smaller datasets, which may not capture the full complexity of acoustic biomarkers and could limit the generalizability of their findings.

To our knowledge, the application of a specialized, pre-trained health acoustic foundation model like HeAR within a multimodal framework on a large-scale dataset for TB detection has not yet been thoroughly investigated. This study aims to fill these gaps by evaluating the performance of HeAR as a powerful feature extractor, combined with clinical data on a large and diverse dataset of 16,000 samples, to potentially set a new benchmark for accuracy and robustness in non-invasive TB screening.

III. Data and Methodology

This section details the data sources and methodological approaches employed to address the core objectives of this research. Fig. 1 illustrates the procedural framework guiding this study.

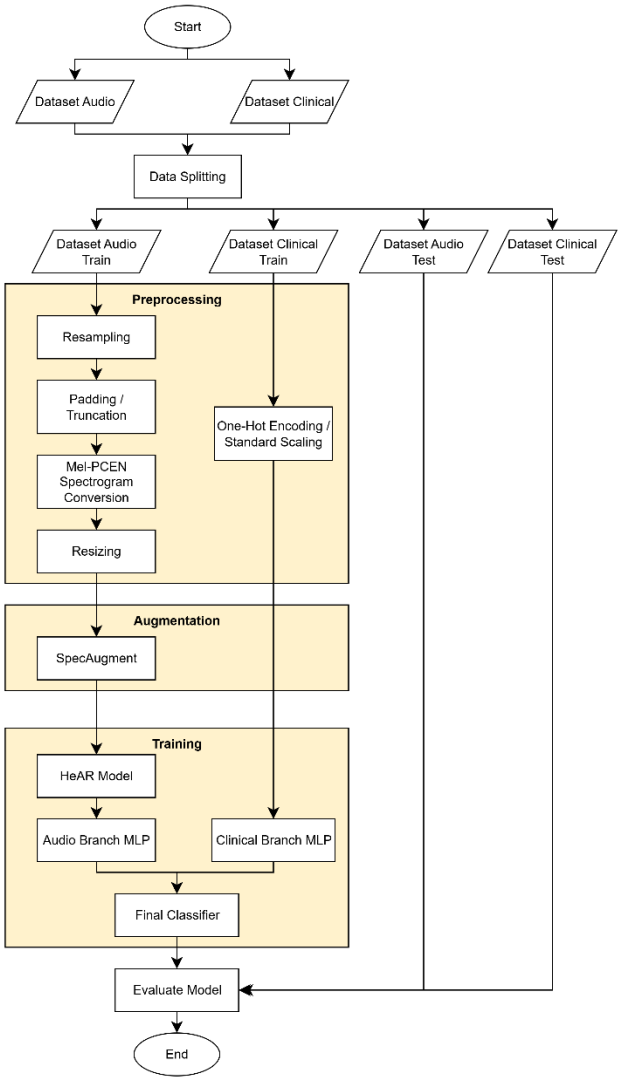


Fig. 1. Research workflow.

A. Dataset

The dataset employed in this research is the CODA TB DREAM Challenge dataset [21], a comprehensive collection of solicited cough recordings designed to advance TB triage testing. The dataset comprises 29,768 cough recordings from individuals across seven countries: India, the Philippines, South Africa, Uganda, Vietnam, Tanzania, and Madagascar.

Each recording is accompanied by detailed metadata, including the participant's confirmed TB status (positive or negative) and a range of demographic and clinical information.

To address the class imbalance often present in medical datasets, we performed random undersampling on the majority class. This process resulted in a balanced final dataset for our study, consisting of 8,000 samples for the TB-positive class and 8,000 samples for the TB-negative class, thereby preventing potential model bias towards the more frequent class. The dataset is illustrated in Fig. 2. While this strategy is effective for balancing classes, we acknowledge that it may discard potentially useful data from the majority class. Alternative techniques, such as the Synthetic Minority Over-sampling Technique (SMOTE) or using focal loss during training, were considered but random undersampling was chosen for its simplicity and computational efficiency in this initial large-scale study. The richness and scale of this balanced 16,000-sample dataset, sourced from seven diverse countries, provide a robust foundation for training a generalizable AI-based screening tool.

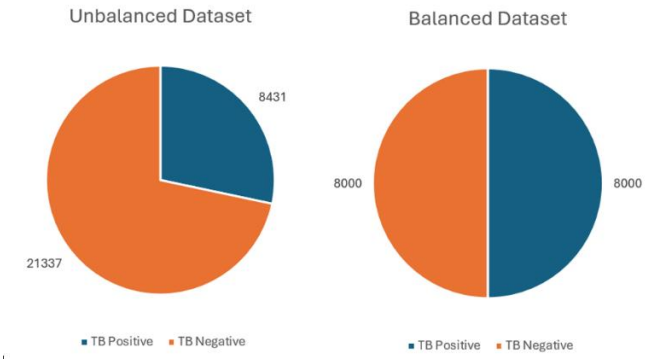


Fig. 2. Dataset distribution.

For this study, the TB status will serve as the primary target label for classification. To develop our multimodal model, we will utilize both the cough audio recordings, and the following supplementary clinical data provided in the metadata: sex, age, and the presence of key TB-related symptoms, namely hemoptysis (coughing up blood), night sweats, weight loss, and fever. The richness and diversity of this dataset make it an ideal resource for training and validating a robust, generalizable AI-based screening tool for tuberculosis.

B. Data Preprocessing and Augmentation

This study applied several data preprocessing techniques, including the following:

- Resampling: This process involves adjusting the sampling rate to match the model's requirements, ensuring input consistency, reducing memory usage, and improving computational efficiency [22]. In this study, we used a sampling rate of 16,000 Hz.
- Padding and Truncation: Padding and truncation are processes used to standardize the duration of audio samples, ensuring that they can be uniformly processed by the model. Padding involves adding zero values to audio samples that are shorter than the target length, whereas truncation trims samples that exceed the specified duration. In this study, the audio length was set to 2 seconds according to the model specification.
- Mel-PCEN Spectrogram Conversion: To create a suitable input representation for the model, the standardized waveforms are converted into a mel-spectrogram format. This process involves applying the Short-Time Fourier Transform (STFT) to decompose

the audio into its frequency components over time, as defined in Eq. (1).

$$X(m,w) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-jwn} \quad (1)$$

where, $x[n]$ is the input signal, $w[n]$ is the window function, and the output $X(m,w)$ is the complex value for the frequency bin w at time frame m. Subsequently, the spectrogram is mapped onto the Mel scale, which more closely aligns with human auditory perception by emphasizing lower frequencies. Following this, Per-Channel Energy Normalization (PCEN) is applied as an advanced dynamic range compression method to enhance signal robustness against noise and volume variations [23]. In this study, the conversion utilized a frame length of 400 samples, a frame step of 160 samples, and generated 128 mel bands.

- Input Resizing: To ensure a consistent input size for the deep learning model, the final mel-PCEN spectrogram was resized to a fixed dimension of 192x128, representing time and frequency axes, respectively.
- SpecAugment: To enhance data diversity and improve the model's generalization capabilities, we applied SpecAugment to the training set. This data augmentation technique operates directly on the spectrogram by randomly applying time masking (obscuring a range of consecutive time steps) and frequency masking (hiding a block of consecutive mel frequency channels). This process encourages the model to learn more robust features and reduces the risk of overfitting [24].

C. Building Model

This study adopts a multimodal deep learning approach that leverages both acoustic features from cough sounds and structured clinical data for TB classification. The architecture consists of two parallel processing streams: one for audio and one for clinical metadata which are then combined through a fusion mechanism before a final classification is made.

The audio processing stream utilizes the pre-trained HeAR model as its feature extraction backbone. The choice of HeAR is deliberate; as a foundation model pre-trained on over 300 million health-related audio clips, it provides powerful and generalized acoustic embeddings, allowing us to achieve high performance without training a deep architecture from scratch. Architecturally, HeAR is a transformer-based model, like a Vision Transformer (ViT), designed to process audio spectrograms as images [25]. The Vision Transformer architecture described in Fig. 3.

The input 192x128 mel-PCEN spectrogram is first divided into a sequence of smaller, non-overlapping patches P_i.

Mathematically, for each patch, the embedding vector E_i is computed using a linear projection, as defined in Eq. (2).

$$E_i = W \cdot vec(P_i) + b \quad (2)$$

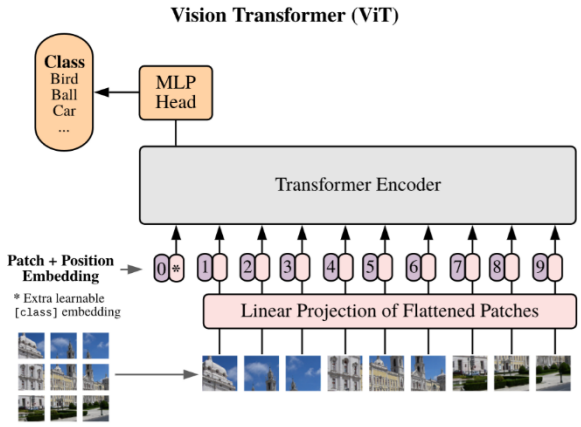


Fig. 3. Vision transformer architecture.

where, W denotes the projection weight matrix, b represents the bias vector, and P_i corresponds to the flattened image patch. To retain spatial information, a learnable positional encoding is then added to each embedding vector. These encodings provide the model with information about the position of each patch in the sequence, using sine and cosine functions of different frequencies, as defined in Eq. (3).

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{100000^{\frac{2i}{d}}}\right)$$
$$PE_{(pos,2i+1)} = \sin\left(\frac{pos}{100000^{\frac{2i}{d}}}\right) \quad (3)$$

where, pos denotes the position index, i represents the dimension index, and d is the embedding dimension.

The resulting sequence of vectors is then processed by 24 transformer encoder layers. Each layer comprises a multi-head self-attention mechanism with 16 attention heads, followed by a position-wise feed-forward network (FFN) with a hidden size of 4096. The self-attention mechanism calculates attention scores based on the query Q, key K, and value V to weigh the importance of different patches relative to each other, allowing the model to capture complex dependencies across the spectrogram as defined in Eq. (4).

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

where, $Q, K,$ and V are the query, key, and value matrices, and d_k is the dimension of the key.

Following the attention sub-layer, the output is passed to an FFN. This network consists of two linear transformations with a GELU activation in between as shown in Eq. (5).

$$FFN(x) = GELU(xW_1 + b_1)W_2 + b_2 \quad (5)$$

where, x denotes the input token representation, W_1 and W_2 are learnable weight matrices, b_1 and b_2 are the corresponding bias vectors, and GELU is the Gaussian Error Linear Unit activation as defined in Eq. (6).

$$GELU(x) = x \cdot \Phi(x) = x \cdot \frac{1}{2}\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}(x + 0.044715x^3)\right)\right) \quad (6)$$

where, $\Phi(x)$ is the standard Gaussian cumulative distribution function. The FFN allows for nonlinear transformations and enhances the model's capacity to learn complex representations. Residual connections are applied around each of the two sub-layers, followed by layer normalization, as defined in Eq. (7). This step is crucial for stabilizing the training of deep networks.

$$Layernorm(x + Sublayer(x)) \quad (7)$$

For this study, the output embedding from the final transformer layer is used as the acoustic feature representations, and the weights of the HeAR backbone are kept frozen. The output which is a 512-dimensional embedding vector is then passed through a dedicated audio branch, which consists of a linear layer that projects the features from 512 to 128 dimensions, followed by Batch Normalization, a ReLU activation, a dropout layer with a rate of 0.4, and a final linear layer that outputs a 64-dimensional audio feature vector.

Concurrently, the clinical data stream handles the tabular metadata. After preprocessing (one-hot encoding for categorical features and standard scaling for age), the vector is fed into a clinical branch. This branch is composed of a linear layer mapping the input to 64 dimensions, followed by a ReLU activation, a dropout layer with a rate of 0.3, and another linear layer that produces a 32-dimensional clinical feature vector.

In the fusion stage, the 64-dimensional audio vector and the 32-dimensional clinical vector are concatenated, creating a combined 96-dimensional feature vector. This vector is then passed to the final classifier which consists of a Batch Normalization layer, a ReLU activation, a high-rate dropout layer (0.5) for regularization, and a final linear layer that outputs the logits for the two classes (TB-positive and TB-negative). The final prediction probabilities are obtained using a softmax function, as shown in Eq. (8).

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad (8)$$

where, z is the input vector of logits and K is the number of classes. The entire model is trained end-to-end using a cross-entropy loss function to quantify the difference between the predicted probability and the actual class labels, as defined in Eq. (9).

$$\mathcal{L} = -\sum_{i=0}^{1} y_i \log(\hat{y}_i) \quad (9)$$

D. Training Setup

The AdamW optimizer was used to train the models, incorporating a weight decay of 0.01 and an initial learning rate of 1×10^{-4}. Training was conducted for 50 epochs with a batch size of 32. The objective function was cross-entropy loss. Early stopping was applied based on validation loss to prevent overfitting, with a patience of 7 epochs and minimum delta of 0.001. Model performance was comprehensively evaluated using several metrics, including accuracy, precision, recall, and F1-score, with 80/10/10 train, validation and test dataset. All experiments were conducted on a NVIDIA RTX A4000 GPU with 16 GB VRAM.

IV. Experimental Results

A. Evaluation Parameters

This study employed several evaluation metrics, including accuracy, F1-score, precision, and recall. The confusion matrix is a fundamental tool for assessing classification model performance, offering a comprehensive comparison between the model's predictions and the actual labels. The confusion matrix comprises four components: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Precision quantifies the proportion of correctly predicted positive instances among all instances predicted as positive, thereby providing insight into the model's susceptibility to false positive errors. Recall, on the other hand, evaluates the proportion of actual positive instances that the model correctly identified, thereby reflecting its effectiveness in minimizing false negative errors. Specificity measures the proportion of actual negatives that are correctly identified as such by the model. The F1-score, computed as the harmonic means of precision and recall, provides a balanced assessment of the model's performance by simultaneously accounting for both FP and FN. These evaluation metrics are computed using the formulas presented in Eq. (10) through Eq. (14).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (10)$$

$$Recall = \frac{TP}{TP + FN} \quad (11)$$

$$Precision = \frac{TP}{TP + FP} \quad (12)$$

$$F1 - Score = \frac{2 \times (precision \times recall)}{precision + recall} \quad (13)$$

$$Specificity = \frac{TN}{TN + FP} \quad (14)$$

B. Experimental Analysis

In this study, simulations and model development for detecting TB based on cough sounds-, as well as the creation of a digital audio-based detection system-, were conducted using hardware and software with specific configurations. Table II presents the detailed specifications of the equipment and tools used.

The authors configured the training process with 50 epochs, a learning rate of 0.0001, and a batch size of 32, using the AdamW optimizer.

As shown in Table III, the proposed multimodal deep learning model demonstrated strong performance in classifying tuberculosis from cough audio and clinical data. The model achieved an overall accuracy of 90% on the test set. The results

show a well-balanced performance across both classes, with precision, recall, and F1-scores of 0.90 for both the "Non-TB" and "TB" classes. This symmetry indicates that the model is equally effective at identifying both positive and negative cases and is not biased towards any single class, which is a crucial attribute for a reliable medical screening tool.

TABLE II. DEVICE SPECIFICATION

Specifications	
GPU	NVIDIA RTX A4000
GPU Memory	16 GB
RAM	16 GB
Disk	1 TB
Programming Language	Python 3.10

TABLE III. MODEL PERFORMANCE

Class	Accuracy	Precision	Recall	F1-Score	Specificity
Non-TB	0.90	0.90	0.90	0.90	0.9
TB	0.90	0.90	0.90	0.90	0.9
Average	**0.90**	**0.90**	**0.90**	**0.90**	**0.9**

The confusion matrix, shown in Fig. 4, provides a more granular view of the model's predictions. Out of 800 non-TB samples, the model correctly identified 723 as Non-TB (True Negatives) and misclassified 77 as TB (False Positives). Similarly, for the 800 TB samples, the model correctly identified 723 as TB (True Positives) and misclassified 77 as Non-TB (False Negatives). The low number of misclassifications for both classes further validates the model's high accuracy and balanced predictive power.

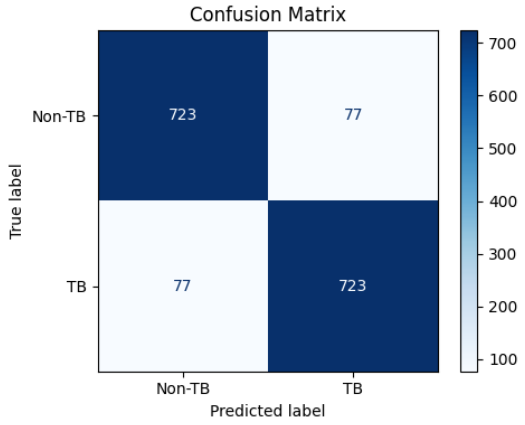


Fig. 4. Confusion matrix of the proposed model on the test set.

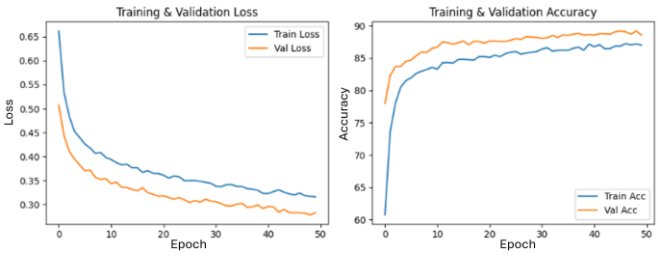


Fig. 5. Training and validation loss and accuracy curves over 50 epochs.

Fig. 5 illustrates the model's learning process over 50 epochs. The training and validation loss curves show a consistent downward trend, indicating that the model was effectively learning from the data. Importantly, the validation loss closely follows the training loss without significant divergence, which suggests that the model generalized well to unseen data and was not overfitting. The accuracy curves mirror this positive trend, with both training and validation accuracy steadily increasing and converging to approximately 90%. This stable training behavior highlights the effectiveness of the chosen architecture and regularization techniques (e.g., dropout) in building a robust classifier.

V. DISCUSSION

Our proposed model achieved a strong overall accuracy of 90%, with balanced performance across all key metrics. This section discusses the interpretation of these findings, the study's limitations, and challenges for real-world deployment.

A. Interpretation of Results and Clinical Implications

Our model's 90% overall accuracy, with balanced precision, recall, and F1-scores of 0.90 across both TB and non-TB classes, demonstrates its reliability and lack of bias. For clinical applications, this performance is highly promising. The 90% sensitivity is crucial for identifying most positive cases early, aiding in timely treatment and reducing transmission. Simultaneously, the 90% specificity effectively rules out the disease in most healthy individuals, minimizing unnecessary follow-up tests, patient anxiety, and healthcare costs. While the 10% false-negative rate requires further work to mitigate public health risks, these findings strongly validate that fusing advanced acoustic features with clinical data offers a holistic and accurate approach for a scalable TB screening tool.

B. Real-World Implementation Scenarios and Challenges

For deployment in real-world TB screening, several challenges must be addressed. Data privacy is paramount, requiring secure data handling. The model's robustness to background noise in uncontrolled environments needs rigorous testing. Furthermore, hardware variability, particularly the quality of microphones in different smartphones, could impact performance and requires investigation to ensure consistent results.

C. Limitations

A primary limitation is our use of random undersampling, which, as noted, may discard potentially useful data. Future studies could address this by implementing alternative balancing techniques like SMOTE or focal loss.

VI. CONCLUSION AND FUTURE WORK

This study successfully demonstrated the effectiveness of a multimodal deep learning approach for tuberculosis detection by combining cough audio signals with clinical metadata. By leveraging the powerful Health Acoustic Representations (HeAR) model as a feature extractor and integrating it with key patient symptoms, our proposed model achieved an impressive accuracy of 90% on a balanced test set. The balanced precision, recall, and F1-scores further underscore the model's capability to reliably identify both Non-TB and TB cases,

highlighting its potential as a robust and unbiased screening tool. These findings confirm that the fusion of advanced acoustic features with contextual clinical data provides a more holistic and accurate representation for diagnosis than either modality alone.

From a clinical perspective, the balance between sensitivity (recall) and specificity is critical for a screening tool. The model's 90% sensitivity indicates that it successfully identifies 9 out of 10 individuals with TB, which is vital for early treatment initiation. However, the 10% of cases missed (false negatives) represent a significant public health concern, as these individuals may unknowingly continue to transmit the disease. Conversely, the 90% specificity means that 9 out of 10 healthy individuals are correctly identified, but the 10% of false positives would be subjected to unnecessary follow-up tests, causing patient anxiety and burdening healthcare resources. For a preliminary screening tool, this performance represents a promising balance, but future work should prioritize further reducing the false-negative rate to maximize the tool's public health impact.

Several avenues for future research can be explored to build upon these promising results. First, extensive hyperparameter tuning of the fusion model, including the architecture of the audio and clinical branches as well as the dropout rates, could further optimize performance. Second, exploring different fusion strategies, such as attention-based mechanisms, may allow the model to dynamically weigh the importance of audio versus clinical features for each sample. Additionally, evaluating the model's generalization on external, unseen datasets from different demographic or geographic populations is crucial to ensure its real-world applicability. Finally, incorporating explainable AI (XAI) techniques is crucial for clinical adoption. By providing insights into which acoustic or clinical features most influence the model's predictions, XAI can build trust among healthcare professionals, aid in validating the model's decision-making process, and facilitate its integration into clinical workflows. The development of such a tool holds significant promise for creating an accessible, low-cost, and scalable solution to support global efforts in early TB detection and containment.

ACKNOWLEDGMENT

The authors like to acknowledge the support of PT Seleris Meditekno Internasional.

REFERENCES

[1] E. C. Jones-L Opez, O. Namugga, F. Mumbowa, M. Ssebidandi, and O. Mbabazi, “Coughing Is Not Required to Transmit Mycobacterium tuberculosis: Another Nail in the Coffin,” https://doi.org/10.1164/rccm.202204-0645ED, vol. 206, no. 2, pp. 141–143, Jul. 2022, doi: 10.1164/RCCM.202204-0645ED.

[2] World Health Organization, “Tuberculosis,” Sep. 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tuberculosis

[3] World Health Organization, “1.1 TB incidence.” Accessed: Sep. 18, 2025. [Online]. Available: https://www.who.int/teams/global-programme-on-tuberculosis-and-lung-health/tb-reports/global-tuberculosis-report-2024/tb-disease-burden/1-1-tb-incidence

[4] M. MacGregor-Fairlie, S. Wilkinson, G. S. Besra, and P. G. Oppenheimer, “Tuberculosis diagnostics: overcoming ancient challenges with modern solutions,” Emerg Top Life Sci, vol. 4, no. 4, p. 435, Dec. 2020, doi: 10.1042/ETLS20200335.

[5] K. K. Chopra and S. Singh, “Tuberculosis: Newer diagnostic tests: Applications and limitations,” Indian Journal of Tuberculosis, vol. 67, no. 4, pp. S86–S90, Dec. 2020, doi: 10.1016/j.ijtb.2020.09.025.

[6] M. G. Moule and J. D. Cirillo, “Mycobacterium tuberculosis Dissemination Plays a Critical Role in Pathogenesis,” Front Cell Infect Microbiol, vol. 10, Feb. 2020, doi: 10.3389/fcimb.2020.00065.

[7] K. C. Rahlwes, B. R. S. Dias, P. C. Campos, S. Alvarez-Arguedas, and M. U. Shiloh, “Pathogenicity and virulence of Mycobacterium tuberculosis,” Virulence, vol. 14, no. 1, Dec. 2023, doi: 10.1080/21505594.2022.2150449.

[8] B. Mathema et al., “Drivers of Tuberculosis Transmission,” J Infect Dis, vol. 216, no. suppl_6, pp. S644–S653, Nov. 2017, doi: 10.1093/infdis/jix354.

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

[10] A. Salma, A. Bustamam, A. Yudantha, A. Victor, and W. Mangunwardoyo, “Artificial Intelligence Approach in Multiclass Diabetic Retinopathy Detection Using Convolutional Neural Network and Attention Mechanism,” International Journal of Advances in Soft Computing and its Applications, vol. 13, no. 3, pp. 101–114, Dec. 2021, doi: 10.15849/IJASCA.211128.08.

[11] A. Bustamam, D. Sarwinda, R. H. Paradisa, A. A. Victor, A. R. Yudantha, and T. Siswantining, “Evaluation of convolutional neural network variants for diagnosis of diabetic retinopathy,” Communications in Mathematical Biology and Neuroscience, 2021, doi: 10.28919/cmbn/5660.

[12] M. Pahar, M. Klopper, B. Reeve, R. Warren, G. Theron, and T. Niesler, “Automatic Cough Classification for Tuberculosis Screening in a Real-World Environment,” Physiol Meas, vol. 42, no. 10, pp. 10.1088/1361-6579/ac2fb8, Oct. 2021, doi: 10.1088/1361-6579/AC2FB8.

[13] R. K. Sahoo, A. Sinha, M. Mishra, D. Bhattacharya, S. Pati, and K. C. Sahoo, “A Systematic Review and Meta-Analysis of the Diagnostic Accuracy of Artificial Intelligence in Detecting Tuberculosis Using Cough Sounds,” 2025, doi: 10.2139/SSRN.5242653.

[14] S. Baur et al., “HeAR-Health Acoustic Representations,” 2024.

[15] A. Ehtesham, A. Singh, S. Kumar, and T. T. Khoei, “Early Detection of Pediatric Pneumonia Using Google’s HeAR Model: A Respiratory Sound Embedding Approach,” 2025 IEEE World AI IoT Congress (AIIoT), pp. 0185–0191, May 2025, doi: 10.1109/AIIOT65859.2025.11105317.

[16] G. P. Kafentzis et al., “Predicting Tuberculosis from Real-World Cough Audio Recordings and Metadata,” Jul. 2023, Accessed: Sep. 19, 2025. [Online]. Available: https://arxiv.org/pdf/2307.04842

[17] S. Huddart et al., “Solicited Cough Sound Analysis for Tuberculosis Triage Testing: The CODA TB DREAM Challenge Dataset,” Mar. 28, 2024. doi: 10.1101/2024.03.27.24304980.

[18] A. Dadu et al., “PASS to End TB in Europe: Accelerated efforts on prevention and systematic screening to end tuberculosis in the WHO European Region by 2030,” International Journal of Infectious Diseases, vol. 141, Apr. 2024, doi: 10.1016/J.IJID.2024.02.023.

[19] G. H. R. Botha et al., “Detection of tuberculosis by automatic cough sound analysis,” Physiol Meas, vol. 39, no. 4, p. 045005, Apr. 2018, doi: 10.1088/1361-6579/AAB6D0.

[20] W. Xu et al., “Feature fusion method for pulmonary tuberculosis patient detection based on cough sound,” PLoS One, vol. 19, no. 5, p. e0302651, May 2024, doi: 10.1371/journal.pone.0302651.

[21] G. P. Kafentzis et al., “Predicting Tuberculosis From Real-World Cough Audio Recordings And Metadata”.

[22] Y. C. Eldar, “Resampling,” Cambridge University Press eBooks, pp. 323–367, Dec. 2014, doi: 10.1017/cbo9780511762321.010.

[23] V. Lostanlen et al., “Per-Channel Energy Normalization: Why and How,” IEEE Signal Process Lett, vol. 26, no. 1, pp. 39–43, Jan. 2019, doi: 10.1109/LSP.2018.2878620.

[24] D. S. Park et al., "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," 2019, doi: 10.21437/Interspeech.2019-2680.

[25] A. Dosovitskiy et al., "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE", Accessed: Sep. 25, 2025. [Online]. Available: https://github.com/

EYE-GDM: Clinically Validated, Explainable Ensemble Learning for Gestational Diabetes

Shatha Alghamdi[1], Rashid Mehmood[2], Fahad Alqurashi[3], Turki Alghamdi[4], Sarah Ghazali[5], Asmaa AlAhmadi[6]
Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia[1,3]
Computer Science & Artificial Intelligence Department, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia[1]
Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, Saudi Arabia[2,4]
Department of obstetrics and gynecology, College of Medicine, King Abdulaziz University, Jeddah, Saudi Arabia[5]
Department of Women and Child Health (Obstetrics and Gynecology), College of Medicine, Taibah University, Madinah, Saudi Arabia[6]

***Abstract*—As artificial intelligence (AI) advances in healthcare, its use in maternal health shows promise but faces challenges of trust due to the black-box nature of many models. Gestational diabetes mellitus (GDM), a transient yet high-risk condition, demands accurate and interpretable prediction tools. However, existing GDM prediction studies often rely on opaque models or post-hoc explanation techniques applied after training, which limits transparency and reduces their clinical applicability. This highlights an urgent need for models that unify high predictive performance with interpretability by design. This study introduces EYE-GDM, a case-specific application of our Enhanced Interpretability Ensemble (EYE) framework, designed to predict GDM risk with clinically meaningful explanations. The pipeline evaluates multiple algorithms and selects Decision Tree (DT), k-Nearest Neighbors (k-NN), and Gradient Boosting (GB) as the best-performing base learners. These are integrated with SHAP and a logistic regression (LR) meta-model to construct EYE-GDM, embedding interpretability by weighting learner outputs with LR coefficients. This yields global (population-level) and local (patient-level) explanations consistent with medical knowledge. Tested on a dataset of 3,525 pregnancies, EYE-GDM achieved strong performance (accuracy = 0.9789, AUC-ROC = 0.9981) and provided insights into risk patterns, thresholds, and feature interactions relevant to GDM. By embedding explainability within the ensemble construction, EYE-GDM achieves transparent and clinically aligned reasoning without compromising predictive performance. Thus, EYE-GDM demonstrates how explainable AI (XAI) can translate from technical innovation to practical value in maternal care, supporting earlier risk identification and more informed clinical decisions.**

***Keywords*—Explainable Artificial Intelligence (XAI); interpretable machine learning (IML); Gestational diabetes mellitus (GDM); maternal health; healthcare AI; GDM risk prediction; transparency; trust**

I. Introduction

Artificial intelligence (AI) is increasingly being used in healthcare to support early risk detection and guide clinical decision-making [1]. Yet, despite its growing presence, many AI models, especially high-performing ones such as deep learning and ensemble methods, suffer from the black-box problem and remain difficult for clinicians to trust and use [2]. In areas such as maternal health, where model predictions can directly affect outcomes for both mother and child, transparency is not optional. Clinicians need more than just a prediction; they need a clear rationale they can understand and explain. In this context, integrating explainable AI (XAI) into decision support systems is essential, as it can significantly affect clinicians' trust and the extent to which they follow AI-driven recommendations [3].

Gestational diabetes mellitus (GDM) affects a substantial number of pregnancies worldwide and is linked to serious complications such as preeclampsia, macrosomia, neonatal hypoglycaemia, and future type 2 diabetes. Its transient nature and rapid physiological onset during pregnancy create a narrow window for timely risk identification and intervention [4]. In addition, GDM has been associated with adverse pregnancy outcomes including preterm birth, hypertensive disorders, shoulder dystocia, hyperbilirubinemia, stillbirth, and caesarean delivery [5], [6], [7]. Therefore, there is an urgent need for clinically interpretable machine learning methods tailored to GDM patients, so that predictions can be understood and applied meaningfully in practice [8].

Although machine learning (ML) has been widely applied to GDM prediction, many models rely on post-hoc explainability methods such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations). These approaches provide some insight but do so after the model has been trained, treating explanation as a separate process. This separation often limits clinical usefulness, as the reasoning behind predictions may not fully reflect the model's internal decision logic [9], [10]; see Section II for related work and research gap.

To address the dual challenge of predictive performance and interpretability, we previously introduced the Enhanced Interpretability Ensemble (EYE) framework [11]. EYE is a structured methodology that embeds interpretability throughout the model development pipeline, from defining the clinical application and preparing data to selecting base learners, integrating explanation methods, and constructing ensembles that remain transparent by design. Unlike post hoc approaches, EYE aligns interpretability with the model's internal reasoning.

We first implemented this framework in EYE-WD [11], focused on diabetes risk prediction in women, where it achieved competitive performance across multiple datasets while uncovering clinically consistent explanations. This work highlighted the value of building AI systems that capture health patterns shaped by gender-specific factors, ensuring that predictive tools serve populations often overlooked in women's and maternal health research.

Building on this foundation, the present work applies the EYE framework to gestational diabetes mellitus (GDM), a transient but high-risk condition unique to pregnancy, resulting in EYE-GDM. The ensemble integrates Decision Tree (DT), k-Nearest Neighbors (k-NN), and Gradient Boosting (GB) as base learners, combined through a logistic regression (LR) meta-model with SHAP explanations embedded in training. Applied to a dataset of 3,525 pregnancies, EYE-GDM achieved an accuracy of 97.89%, F1-score of 97.90%, precision of 97.47%, recall of 98.33%, and an AUC-ROC of 0.9981. Beyond these results, interpretability analyses produced clinically meaningful insights. Global SHAP patterns revealed thresholds such as elevated risk at BMI values above 25 and OGTT values exceeding 160 mg/dL, while HDL levels above 35 mg/dL were protective. Interactions between features, for example, high BMI with low HDL or multiple pregnancies with elevated blood pressure, revealed clusters of maternal risk factors. Local explanations further traced patient-specific outcomes, distinguishing protective from adverse profiles in a transparent way.

In doing so, this study advances GDM risk assessment and supports the broader goal of developing XAI that can be integrated into real-world maternal-care workflows [12]. By presenting EYE-GDM alongside the earlier EYE-WD implementation, we demonstrate both the flexibility and clinical relevance of the EYE framework and its potential to address critical challenges in women's and maternal health, providing contributions that are meaningful to technical experts and healthcare professionals alike.

The remainder of this study is organized as follows: Section II reviews the related work. Section III provides the methodology. Section IV reports the performance results of EYE-GDM, while Section V provides clinically oriented interpretations, including dataset-specific risk patterns, feature interactions, data-driven thresholds, and patient-level interpretations. Section VI provides a discussion, and Section VII concludes the study with directions for future work.

II. Related Work

GDM has received growing attention from both clinical and computational research communities. In particular, a substantial body of work has applied machine learning techniques to develop predictive models for GDM, aiming to support early diagnosis and timely intervention that can improve maternal and neonatal outcomes. Section II A reviews studies that develop machine learning and deep learning models to identify women at risk of GDM. Section II.B examines a growing body of work on making these predictions understandable to clinicians.

A. Machine Learning and Deep Learning for GDM

Sumathi et al. [13], [14] constructed a GDM dataset, which includes 3,525 pregnant women's data. Using deep learning, they introduced a deep stacked autoencoder based on outlier detection process (OD-DSAE), achieving 96.18% accuracy [13], and later proposed an ensemble voting strategy combining k-nearest neighbours (k-NN), Random Forest (RF), and LR models, reaching 94.24% accuracy [14]. Jader et al. [15] employed an ensemble approach using lab records from Iraq-Kurdistan, achieving 92% accuracy through majority voting with Decision Tree (DT), RF, Support Vector Machine (SVM), k-NN, LR, and Naïve Bayes (NB). Gallardo-Rincón et al. [16] developed and evaluated an Artificial Neural Network (ANN) model, achieving 70.3% accuracy in recognizing women at high risk of developing GDM. This model was based on data from 1,709 pregnant Mexican women who participated in the 'Cuido mi embarazo' study. Zheng et al. [17] collected data on 4,771 Chinese pregnant women from Xinhua Hospital. They applied Multivariate LR using Bayesian inference and achieved an accuracy of 64% and an AUC of 0.766. Shen et al. [18] evaluated several machine learning algorithms, including LR, SVM, RF, AdaBoost, DT, NB, k-NN, XGBoost, and Gradient Boosting Decision Tree (GBDT), to predict GDM in areas with limited resources. The algorithms were trained on 12,304 cases from the First Affiliated Hospital of Jinan University and were validated on 1,655 pregnant patients at the Prince of Wales Hospital. At 0.78 AUC, SVM showed the best performance.

Wu et al. [19] et al. utilized LR, k-NN, SVM, and deep neural network (DNN) for early GDM prediction. They evaluated these machine learning models using data from the International Peace Maternal and Child Health Hospital at Shanghai Jiao Tong University School of Medicine, which included 32,190 pregnant patients. DNN obtained the best AUC of 0.8, followed by LR with an AUC of 0.77. For predicting GDM, Hu et al. [20] utilized and compared between traditional LR method and the XGBoost machine learning model on a total of 925 pregnant women. They demonstrated that XGBoost outperformed LR with an AUC of 0.946 and an accuracy of 87.5%. Through the Japan Environment and Children's Study (JECS), Watanabe et al. [21] gathered GDM data from 82,698 expectant women. They employed LR, SVM, GBDT, and RF for gestational diabetes detection. They demonstrated that GBDT produced the best AUC of 0.74.

Table I summarizes machine learning and deep learning research used to predict women at risk of GDM. The first two columns provide the reference number and a summary of each work. The third column presents the reported performance, while the fourth column records limitations and key observations.

TABLE I. SUMMARY OF WORKS ON GDM PREDICTION

Ref	Work Summary	Performance	Limitations / Observations
[13]	Developed a deep learning model using a stacked autoencoder with outlier detection (OD-DSAE).	Accuracy: 96.18%	• Lack interpretability
[14]	Constructed an ensemble voting model using k-NN, RF, and LR.	Accuracy: 94.24%	
[15]	Trained and evaluated an ensemble of six classifiers (DT, RF, SVM, k-NN, LR, NB) on data from Iraq-Kurdistan using majority voting.	Accuracy: 92%	
[16]	Evaluated the ANN model on data from 1,709 pregnant Mexican women.	Accuracy: 70.3%	• Low performance • Lack interpretability
[17]	Applied Multivariate LR using Bayesian inference on 4,771 Chinese pregnant women from Xinhua Hospital.	Accuracy: 64%, AUC: 0.766	
[18]	Compared nine ML models (LR, SVM, RF, AdaBoost, DT, NB, k-NN, XGBoost, and GBDT) trained on 12,304 cases and were validated on 1,655 pregnant patients.	AUC: 0.78 (SVM best)	
[19]	Evaluated LR, k-NN, SVM, and DNN on the Shanghai dataset (32,190 pregnant patients).	AUC: 0.8 (DNN best)	• Moderate performance • Lack interpretability
[20]	Compared LR and XGBoost on 925 pregnant women.	Accuracy: 87.5% (XGB best)	
[21]	Assessed a large-scale Japanese cohort (82,698 pregnant women) with four models LR, SVM, GBDT, and RF.	AUC: 0.74 (GBDT best)	• Low performance • Lack interpretability

B. Explainability and Interpretability in GDM Prediction

Despite these advances in predictive performance, most previous works did not incorporate interpretability into their predictive models. To the best of our knowledge, only a limited number of studies have explicitly integrated explainability. For instance, Du et al. [22] applied SHAP with multiple machine learning models (AdaBoost, LR, SVM, RF, XGBoost) and used Synthetic Minority Oversampling (SMOTE) to balance the data. They reported an accuracy of 76.1%, with SVM performing the best. Khanna et al. [23] evaluated several machine learning pipelines and data balancing strategies for GDM prediction, where a stacking ensemble trained on SMOTE-ENN (Synthetic Minority Oversampling using Edited Nearest Neighbour) data achieved the best performance with an accuracy of 96% and an AUC of 0.96. They applied five post-hoc methods, including SHAP, LIME, ELI5, Qlattice, and Anchor, which confirmed the links between visceral fat levels, the child's birth weight, and GDM.

Zaky et al. [24] explored a range of algorithms, from traditional models such as LR and DT to advanced ensembles such as RF, Gradient Boosting (GB), CatBoost, XGBoost, and LightGBM, and then combined them in a stacking framework with LR as the meta-classifier. The stacked model trained on 26 selected features gave the best results with an accuracy of 88.8%, outperforming all individual models. SHAP was applied as a post-hoc tool to interpret predictions and highlight key biomarkers. While these contributions have advanced GDM prediction, they largely treat explainability as an afterthought rather than a property embedded within the model itself.

Table II summarizes prior works on GDM prediction that incorporate explainability or interpretability. Column 1 lists the reference; Column 2 summarizes the work; Column 3 reports performance metrics; Column 4 notes limitations and observations.

C. Research Gap

Nevertheless, the studies that do incorporate explainability rely mainly on post-hoc methods such as SHAP and LIME, applied only after the model has been trained. While these approaches offer valuable insights, they do not influence the model's internal structure or learning process, so interpretability remains an add-on rather than a built-in property.

Our overarching aim is to deliver accurate and transparent decision support in high-stakes domains by embedding interpretability across the full modelling workflow, aligning design and evaluation with stakeholder requirements to strengthen accountability, and enabling real-world use without sacrificing predictive performance. We previously introduced the FIXAIH framework [25], which articulates six well-defined design priorities to guide the development of trustworthy, interpretable AI for healthcare. Within this vision, we also proposed the EYE framework [11], which operationalizes selected FIXAIH priorities into a concrete architecture, providing interpretability by design, clinical validation, and transparent model development. We initially implemented EYE for diabetes risk in women (EYE-WD).

In this work, we extend EYE to GDM and refer to this implementation as EYE-GDM. Our approach integrates SHAP within the training pipeline, combining base-model outputs via meta-model weights to produce both local and global explanations that reflect the model's internal structure and how predictions are formed, while maintaining strong predictive performance by employing top-performing base learners during model construction. This work improves GDM risk assessment while contributing to the wider aim of building XAI systems suitable for integration into routine maternal-care workflows.

TABLE II. SUMMARY OF RELATED WORKS ON GDM PREDICTION INCORPORATING EXPLAINABILITY

Ref	Work Summary	Performance	Limitations / Observations
[22]	Applied SHAP with multiple ML models: AdaBoost, LR, SVM, RF, and XGBoost with SVM performing the best.	Accuracy: 76.1%	Post-hoc explainability only; not embedded
[23]	Evaluated multiple ML pipelines and data balancing strategies. A stacking ensemble trained on SMOTE-ENN gave the best results. SHAP, LIME, ELI5, Qlattice, and Anchor were applied to interpret results.	Accuracy: 96%, AUC: 0.96	
[24]	Explored a wide range of ML models (LR, DT, NB, RF, GB, CatBoost, XGBoost, LIGHTGBM), combined in a stacking framework with LR as meta-classifier. SHAP was used to explain predictions and rank biomarkers.	Accuracy: 88.8%	
EYE-GDM	Constructed an interpretable ensemble using EYE framework, integrating top-performing base learners, SHAP tool, and LR as meta-model. SHAP is applied internally during training, combining base learner outputs via meta-model weights to generate explanations.	Accuracy: 97.89%, AUC: 0.99	• Implementation of EYE framework • Embedded interpretability • Competitive performance

III. METHODOLOGY AND DESIGN

In this section, we present the methodology and design of the EYE-GDM system. Owing to space limitations, only a summary is provided here. A detailed description of the methodology, the EYE framework, and its initial implementation in EYE-WD can be found in [11]. EYE-GDM builds upon this foundation by extending the EYE framework into a domain-specific configuration tailored for gestational diabetes. The methodological refinements include optimizing the selection of ensemble base learners for pregnancy-related data characteristics, applying explanation-weighted meta-learning to integrate model reasoning with prediction, and aligning interpretability evaluation with pregnancy-specific clinical variables and thresholds. These adaptations expand the framework's methodological scope while preserving its interpretability-by-design foundation. Fig. 1 shows the architecture of the EYE-GDM system, which consists of eight components.

The Application Selection (APS) targets GDM risk prediction. In Data Preparation and Management (DPM), we use a publicly available dataset of 3,525 patients described by 15 clinical and demographic features, labelled into two classes: 2,153 non-GDM and 1,372 GDM cases. The dataset was collected by researchers at SASTRA Deemed to be University, Kumbakonam, Tamil Nadu, India, in collaboration with a consultant obstetrician and gynaecologist from Anbu Hospital and Nalam Clinic. It is publicly available on Kaggle under the CC BY-NC-SA 4.0 license [26] and was previously used in [13] and [14] to develop and evaluate GDM prediction models. The features are Age, Body Mass Index (BMI), Hemoglobin, Systolic Blood Pressure (Sys BP), Diastolic Blood Pressure (Dia BP), Oral Glucose Tolerance Test (OGTT), Prediabetes, Polycystic Ovary Syndrome (PCOS), Family History, Sedentary Lifestyle, Number of Pregnancies, Large Child or Birth Defect, Gestation in Previous Pregnancy, Unexplained Prenatal Loss, and High-Density Lipoprotein (HDL) (see Table III). All features are continuous or binary (0/1); binaries use 0 for absence and 1 for presence. We apply class-wise imputation for missing values, treat outliers using interquartile range (IQR) rules, scale features, and create a stratified train-test split to preserve class distribution.

Base-Model Evaluation and Selection (BMES) evaluates LR, Ridge Classifier (RC), DT, k-NN, NB, RF, AdaBoost, GB, Histogram Gradient Boost (HGB), and XGBoost on the training split using stratified k-fold cross-validation with hyperparameter tuning. We record AUC-ROC, F1-score, precision, recall, and accuracy, then select the top-performing models as the base learners for EYE-GDM.

Interpretable Meta-Model Selection (IM2S) uses LR as the stack combiner because it provides readable coefficients, well-calibrated probability scores, and efficient training and inference [27]. Explainability Tool Selection (XTS) adopts SHAP since it is model-agnostic, grounded in game-theoretic principles, supports dataset-level and patient-level views, and explains how features move a prediction from a baseline [28], [29].

TABLE III. DESCRIPTION OF FEATURES IN GDM DATASET

Feature	Description
Age	Maternal age at time of pregnancy
BMI	Body Mass Index
Hemoglobin	Hemoglobin level
Sys BP	Systolic Blood Pressure
Dia BP	Diastolic Blood Pressure
OGTT	Oral Glucose Tolerance Test result
Prediabetes	History of prediabetes
PCOS	Polycystic Ovary Syndrome diagnosis
Family History	Family history of diabetes
Sedentary Lifestyle	Lack of regular physical activity
No of Pregnancy	Number of prior pregnancies
Large Child or Birth Default	Previous large baby or complicated delivery
Gestation in Previous Pregnancy	Prior gestational diabetes or complications
Unexplained prenatal loss	History of miscarriage or fetal loss without known cause
HDL	High-Density Lipoprotein cholesterol

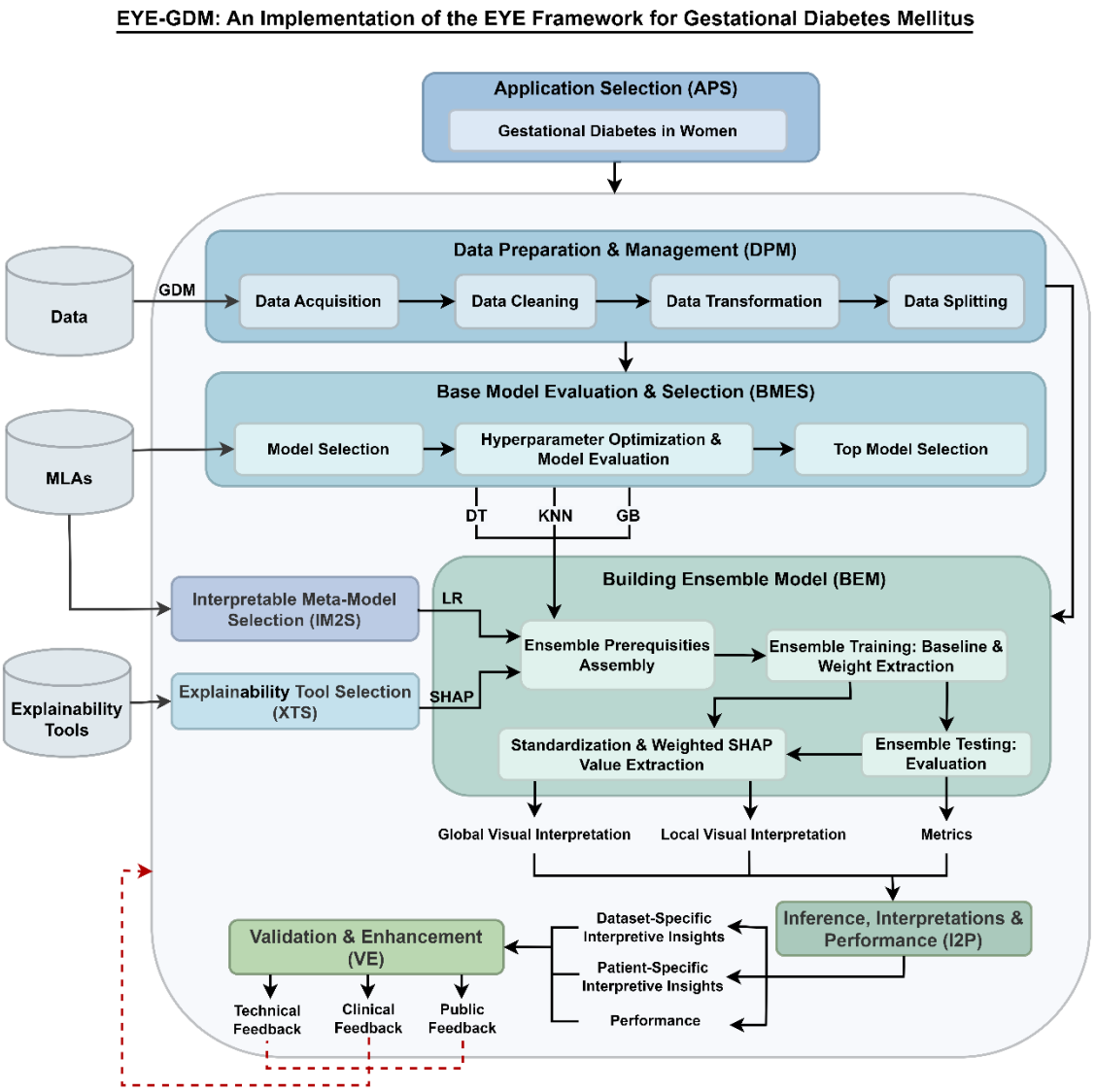


Fig. 1. The EYE-GDM System: An implementation of the EYE framework for gestational diabetes.

Building the Ensemble (BEM) constructs the interpretable ensemble. During training, each base learner yields SHAP values and expected values. We standardize these values to make magnitudes comparable, then form a single explanation by weighing each model's SHAP vector with the learned LR coefficients. This produces consolidated local and global explanations aligned with the ensemble's decision rule. Within BEM, these explanations are rendered as SHAP summary plots (global interpretations), dependence plots (pairwise effects and thresholds), and waterfall plots (patient-level breakdown).

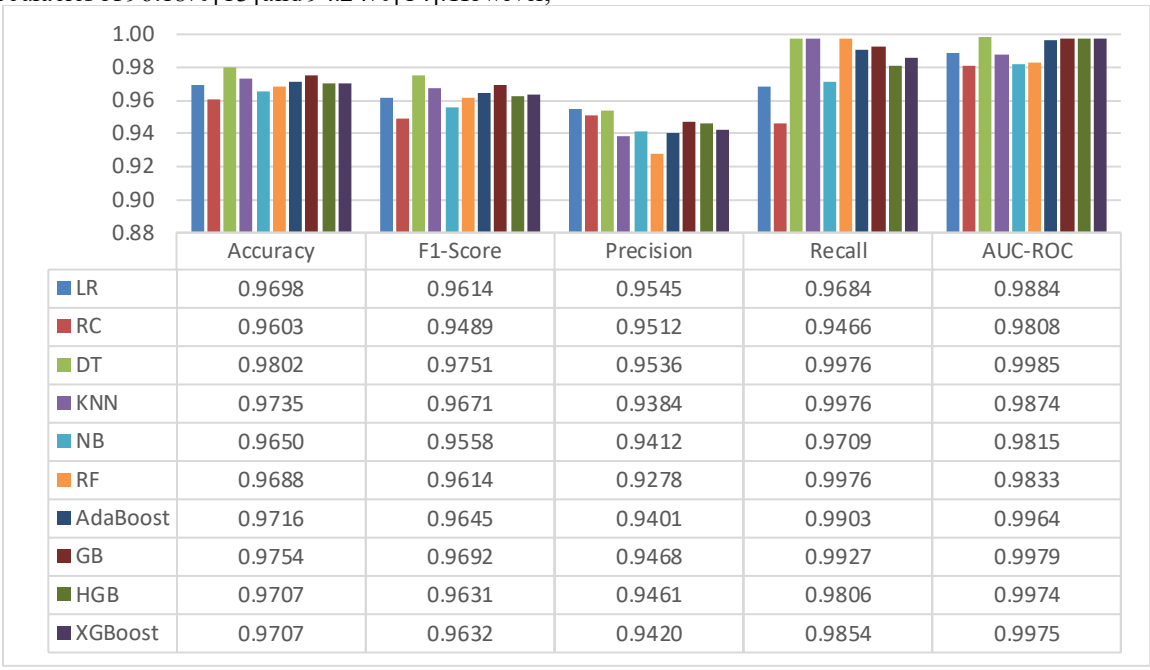
Inference, Interpretation, and Performance (I2P) uses these plots and prediction scores to deliver clinically relevant, fused explanations at both patient and dataset levels, clarifying which factors raise or lower risk, identifying data-driven thresholds, and characterizing feature interactions. Validation and Enhancement (VE) validates both technical and clinical outcomes: technical validation benchmarks performance against prior studies and includes review by technical experts, while clinical validation is conducted by clinical experts to assess alignment with medical standards. Their evaluation focused on adherence to medical knowledge, consistency with clinical practice, validity of diagnostic reasoning, clarity of explanations, and recommendations for improvement. Accepted adjustments are then applied and re-tested on the held-out split.

IV. EYE-GDM: Results

In this section, we present results from the BMES component, which evaluates multiple ML algorithms to identify the top-performing base learners for constructing EYE-GDM. Fig. 2 shows performance across five metrics (accuracy, F1, precision, recall, and AUC-ROC). DT, k-NN, and GB achieved the highest accuracies, with AUC-ROC values around 0.98, indicating clear separation between diabetic and non-diabetic cases. These three models were therefore selected as the ensemble's base learners. During the training phase, the LR meta-model determined contribution weights for each base

learner according to its influence on the overall ensemble output. The resulting coefficients were 6.96 for DT, 1.86 for k-NN, and 0.77 for GB, reflecting their relative influence in the final prediction. These weights were then applied to combine the base learners' explanation outputs, ensuring that the ensemble's interpretability aligns with its internal decision logic.

We evaluated EYE-GDM on the GDM dataset and compared its predictive performance with prior works that used the same data. As shown in Table IV, the reference models achieved accuracies of 96.18% [13] and 94.24% [14]. However, those studies reported only performance and did not provide interpretability or clinical explanations. In contrast, EYE-GDM achieved an accuracy of 97.89%, an F1-score of 97.90%, a precision of 97.47%, a recall of 98.33%, and an AUC-ROC of 0.9981. These results highlight that the framework shows competitive performance compared to previously reported baselines in predictive reliability. More importantly, it does so while also offering transparent and clinically meaningful explanations of its predictions, which prior models lacked. This dual capability of competitive performance and interpretability forms the basis for the subsequent analysis of GDM risk factors.



	Accuracy	F1-Score	Precision	Recall	AUC-ROC
LR	0.9698	0.9614	0.9545	0.9684	0.9884
RC	0.9603	0.9489	0.9512	0.9466	0.9808
DT	0.9802	0.9751	0.9536	0.9976	0.9985
KNN	0.9735	0.9671	0.9384	0.9976	0.9874
NB	0.9650	0.9558	0.9412	0.9709	0.9815
RF	0.9688	0.9614	0.9278	0.9976	0.9833
AdaBoost	0.9716	0.9645	0.9401	0.9903	0.9964
GB	0.9754	0.9692	0.9468	0.9927	0.9979
HGB	0.9707	0.9631	0.9461	0.9806	0.9974
XGBoost	0.9707	0.9632	0.9420	0.9854	0.9975

Fig. 2. Performance comparison across candidate base learners.

TABLE IV. PERFORMANCE COMPARISON OF EYE-GDM AGAINST PRIOR STUDIES ON THE SAME DATASET

Work	Accuracy	F1-score	Precision	Recall	AUC-ROC
[13]	0.9618	0.9741	0.9617	0.9869	-
[14]	0.9424	0.9400	0.9400	0.9400	-
EYE-GDM	0.9789	0.9790	0.9747	0.9833	0.9981

V. EYE-GDM: CLINICAL INTERPRETATIONS

In this section, we provide interpretations at both the dataset-specific and the individual patient-specific, alongside an analysis of feature interactions and dataset-specific thresholds. Different datasets, representing distinct populations, may yield unique thresholds influenced by lifestyle and other population-specific factors [30]. Therefore, the thresholds identified here should be regarded as dataset-specific, and validation across different populations is necessary. In Section V A, we use a SHAP summary plot to interpret global feature contributions to GDM risk prediction, thereby identifying dataset-specific interpretations. SHAP dependence plots are then applied in Section V B to capture feature interactions and define dataset-specific thresholds. Finally, in Section V C, we illustrate patient-level interpretation by selecting a representative patient case and applying a SHAP waterfall plot. In these SHAP plots, positive SHAP values indicate a higher likelihood of GDM, whereas negative SHAP values reflect a lower predicted risk.

A. Dataset-Specific Interpretations

The SHAP summary plot (Fig. 3) illustrates the global feature contributions to GDM risk predictions across the dataset. It offers a compact view of how feature values shift the prediction toward or away from GDM for all patients. The y-axis lists features, while the x-axis shows SHAP values, which represent a feature's impact on the model output for each case: positive values push the prediction toward GDM, negative values toward non-GDM, and values farther from zero indicate stronger effects. Each dot corresponds to one patient's contribution for that feature. Dot color encodes the raw feature value for that patient, with blue indicating low values and red indicating high values as shown in the legend. When red points cluster on the positive side, higher feature values are associated with higher predicted risk; when red points cluster on the negative side, higher values are associated with lower predicted risk; mixed colors on both sides suggest non-linear or context-dependent effects.

From a general health perspective, higher BMI continues to exhibit a strong and consistent positive SHAP trend, confirming that elevated adiposity contributes directly to higher GDM risk. In contrast, maternal age presents a less uniform pattern: while younger age values tend to cluster around zero and negative SHAP values (indicating lower predicted risk), older age values are associated with a wider SHAP range, suggesting that age may either increase or decrease risk depending on the presence of co-occurring factors such as BMI or blood pressure. Higher Hemoglobin levels are associated with increased predicted GDM risk, whereas moderate and lower levels generally correspond to lower SHAP values, indicating reduced model-predicted risk.

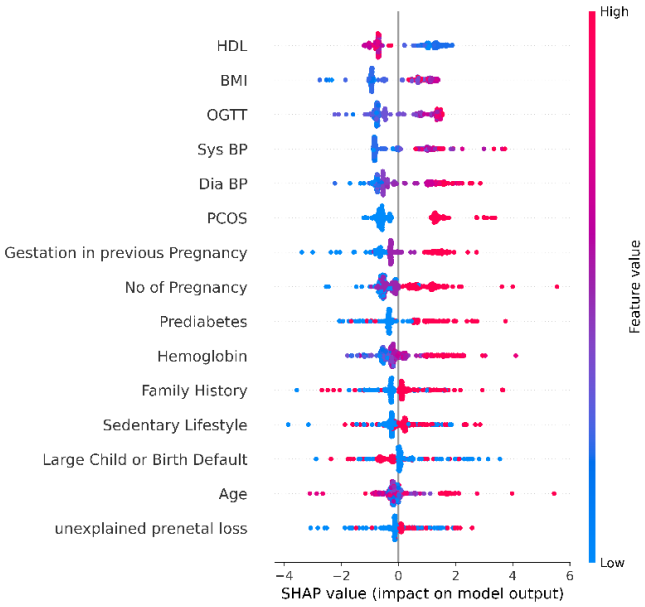


Fig. 3. SHAP summary plot.

For cardiovascular status, both Sys BP and Dia BP display positive SHAP values at elevated readings, indicating that hypertensive disorders in pregnancy often co-occur with insulin resistance and elevated GDM risk. From a metabolic and endocrine perspective, elevated OGTT values substantially increase the predicted risk of GDM. Similarly, the presence of PCOS contributes strong positive SHAP values, indicating higher risk. In contrast, HDL shows a clear inverse association, where lower levels correspond to an increased likelihood of GDM.

Features related to obstetric and reproductive history include Gestation in Previous Pregnancy and Number of Pregnancies, both of which tend to show positive contributions at higher values. A positive history of Unexplained Prenatal Loss is also associated with increased predicted risk. In contrast, Large Child or Birth Default exhibits a more ambiguous pattern, with both its presence and absence producing mixed SHAP contributions. This suggests a feature interaction, where the model may shift predictive importance to other highly correlated variables, such as BMI, maternal age, or parity, thereby diminishing the apparent effect of this obstetric factor.

Finally, Prediabetes, Family History and Sedentary Lifestyle show SHAP patterns where their presence typically contributes to a higher predicted risk of GDM. However, in a small number of cases, even when these risk factors are present, they are linked to lower SHAP values (lower GDM risk), likely due to the dominant influence of other protective factors or interactions that offset their individual impact.

B. Factor Interactions and Dataset-Specific Thresholds

This subsection highlights how clinical and obstetric factors interact to shape the model's GDM risk predictions. SHAP dependence plots (Fig. 4 to Fig. 18) show how the effect of one feature varies with another's value and surface the strongest interactions, helping to spot practical thresholds and non-linear links useful for clinical review. In each plot, the x-axis is the primary feature value, the y-axis is its SHAP value (its contribution to predicted risk), each point is a single case, and a blue-to-red color scale encodes the value of a second interacting feature. Read together, these plots clarify combined patterns such as high BMI with low HDL, or multiple pregnancies with elevated blood pressure, that contribute to higher predicted risk and are expressed in familiar clinical terms to support decision-making.

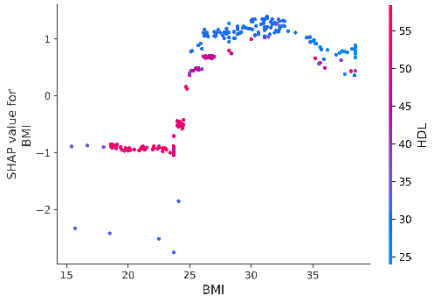


Fig. 4. SHAP dependence plot (BMI & HDL).

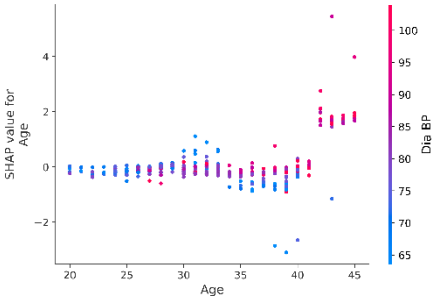


Fig. 5. SHAP dependence plot (Age & Dia BP).

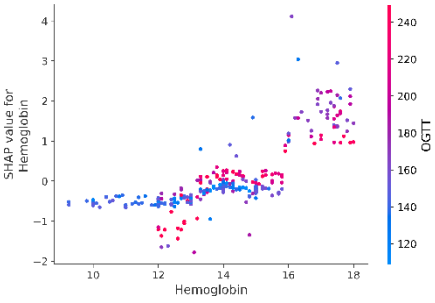


Fig. 6. SHAP dependence plot (Hemoglobin & OGTT).

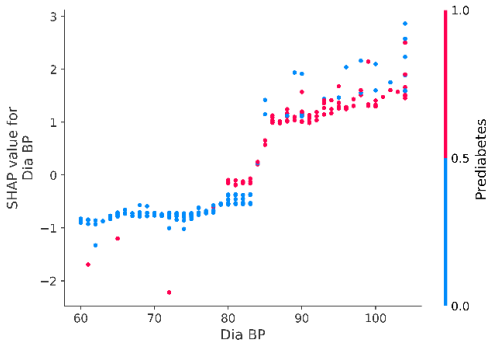


Fig. 7. SHAP dependence plot (Dia BP & Prediabetes).

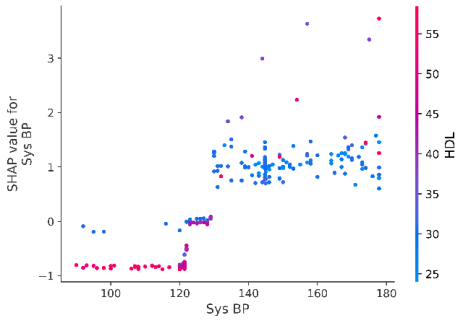


Fig. 8. SHAP dependence plot (Sys BP & HDL).

In Fig. 4, BMI shows a clear threshold around 25. Below this value, SHAP contributions are typically negative, often accompanied by high HDL levels, indicating lower risk. Beyond 25, SHAP values rise sharply, especially in individuals with low HDL. For Maternal Age (Fig. 5), younger women tend to have neutral or slightly negative SHAP values, whereas women above 40 years consistently show positive contributions and are associated with elevated Dia BP. In Fig. 6, Hemoglobin contributes positively to GDM risk between 13 and 16 g/dL, but only when OGTT values are elevated, whereas lower OGTT levels are linked to reduced risk. At levels above 16 g/dL, Hemoglobin consistently shows positive contributions to GDM risk and is associated with high OGTT values. This pattern is important and may be explained by physiological changes such as increased blood viscosity [31], indicating that both factors are likely working together rather than high Hemoglobin level alone.

In the cardiovascular domain, both Dia BP and Sys BP exhibit clear threshold behaviour. Fig. 7 illustrates that Dia BP values above 85 mmHg are associated with positive SHAP contributions, particularly in individuals with prediabetes, whereas values below 80 mmHg in the absence of prediabetes indicate lower risk. Similarly, Fig. 8 shows that Sys BP contributes to higher risk beyond 130 mmHg, with stronger effects observed in patients with low HDL. In contrast, values below 120 mmHg combined with high HDL levels are generally linked to lower risk.

Among metabolic and endocrine features, Fig. 9 shows that OGTT values rise sharply in SHAP contributions beyond 160 mg/dL, particularly in patients with high Sys BP, whereas values below 150 mg/dL are considered protective. In Fig. 10, PCOS consistently contributes to higher predicted risk when present, with its effect being stronger in individuals with high BMI. Fig. 11 illustrates that HDL values below 35 mg/dL are associated with increased risk, especially in women with prior gestations, while values above 35 mg/dL correspond to protective SHAP contributions. This pattern suggests that HDL remains a reliable marker of metabolic health, regardless of gestational history.

Obstetric and reproductive history features show interpretable patterns. Fig. 12 illustrates that Gestation in Previous Pregnancy contributes positively when equal to or greater than two, particularly in patients with elevated Dia BP. Fig. 13 shows that Number of Pregnancies follows a similar upward trend, with risk increasing beyond three. Fig. 14 indicates that Unexplained Prenatal Loss, when present, raises the probability of GDM, although its absence does not necessarily imply low risk. Fig. 15 demonstrates that Prediabetes, when combined with high BMI, is linked to higher GDM risk, whereas its absence together with low BMI is associated with reduced risk.

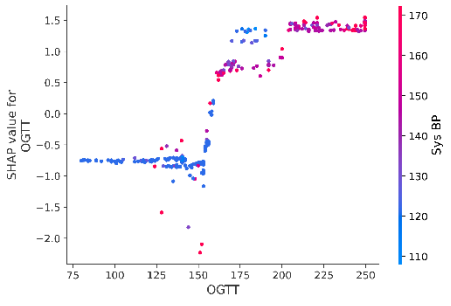


Fig. 9. SHAP dependence plot (OGTT & Sys BP).

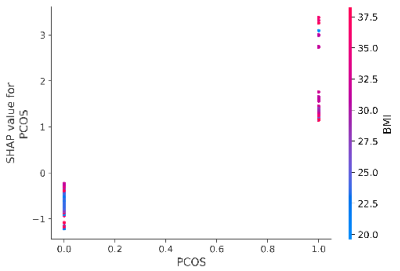


Fig. 10. SHAP dependence plot (PCOS & BMI).

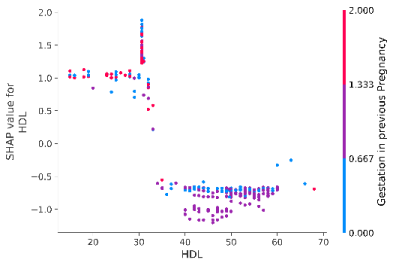


Fig. 11. SHAP dependence plot (HDL & gestation in previous pregnancy).

In contrast, Fig. 16 to Fig. 18 (Large Child or Birth Default, Family History, and Sedentary Lifestyle) show less consistent behavior. Their SHAP contributions vary widely, suggesting that their predictive influence may be partly absorbed by stronger correlated factors.

Collectively, these dependence plots reveal that individual factors influence risk through threshold effects and interaction with other variables. They reinforce clinical understanding of GDM risk while offering finer-grained, data-driven insight into how multiple features combine to shape the model's predictions. Table V summarizes these insights by aligning the data-derived thresholds from SHAP analysis with established clinical reference ranges, highlighting key feature interactions and their associated GDM risk. This comparative view helps contextualize model behavior against known medical standards, reinforcing the interpretability and clinical relevance of EYE-GDM results.

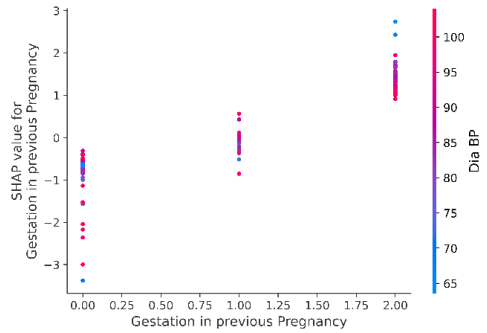


Fig. 12. SHAP dependence plot (gestation in previous pregnancy & Dia BP).

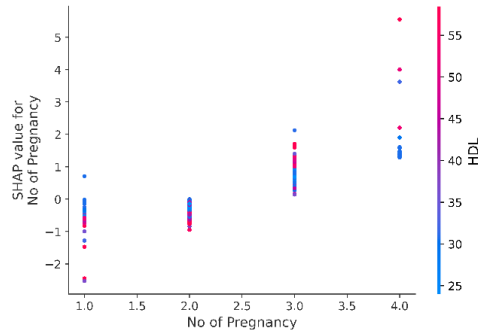


Fig. 13. SHAP dependence plot (number of pregnancy & HDL).

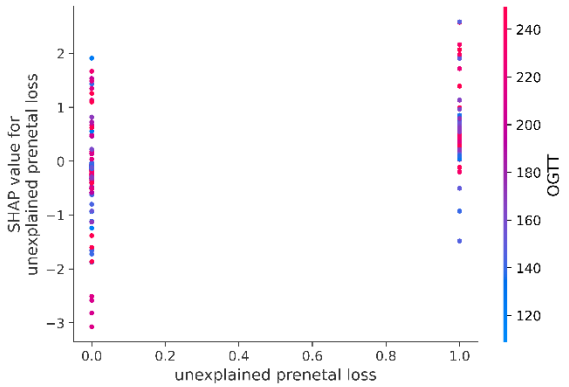


Fig. 14. SHAP dependence plot (unexplained prenatal loss & OGTT).

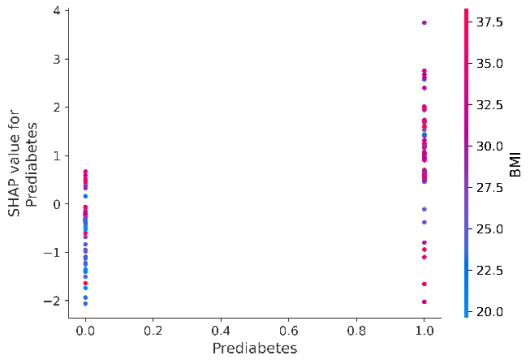


Fig. 15. SHAP dependence plot (prediabetes & BMI).

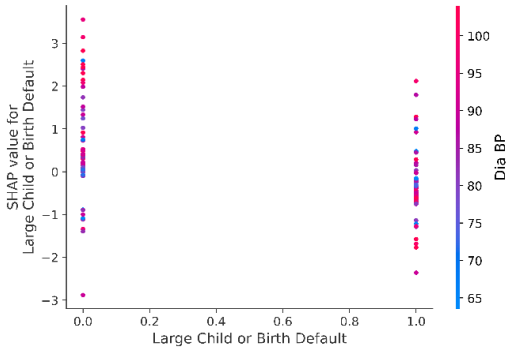


Fig. 16. SHAP dependence plot (large child or birth default & Dia BP).

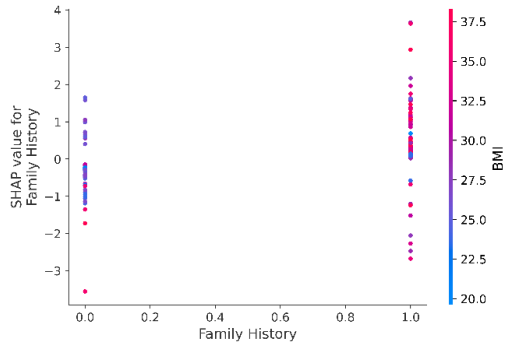


Fig. 17. SHAP dependence plot (family history & BMI).

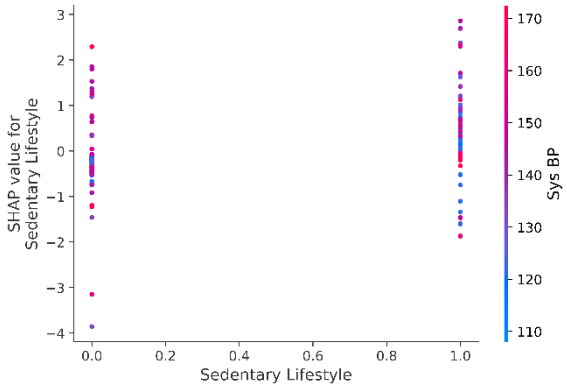


Fig. 18. SHAP dependence plot (sedentary lifestyle & sys BP).

TABLE V. DATASET-DERIVED THRESHOLDS, INTERACTING FACTORS, AND CLINICAL NORMAL RANGES FOR KEY GDM PREDICTORS

Feature	Clinical Cutoff / Normal Range	Data-driven Threshold	Interacts with	GDM Risk
BMI	18.5–24.9 kg/m² [32], [33]	> 25	Low HDL	High
		< 25	High HDL	Low
Age	-	> 40 years	High Dia BP	High
Hemoglobin	12-16 g/dL [34]	13-16 g/dL	High OGTT	
		> 16 g/dL	-	
Dia BP	< 80 mmHg (normal), 80-89 mmHg (Stage 1 hypertension), ≥ 90 mmHg (Stage 2 hypertension) [35]	> 85 mmHg	Prediabetes Presence	
		< 80 mmHg	Prediabetes Absence	Low
Sys BP	< 120 mmHg (normal), 120-129 mmHg (elevated), 130-139 mmHg (Stage 1 hypertension), ≥ 140 mmHg (Stage 2 hypertension) [35]	> 130 mmHg	Low HDL	High
		< 120 mmHg	High HDL	Low
OGTT	75-g 2-h: <153 mg/dL [32]	> 160 mg/dL	High Sys BP	High
		< 150 mg/dL	-	Low
PCOS	-	Presence	High BMI	High
HDL	50-80 mg/dL (women) [36]	< 35 mg/dL	Gestational History	
		> 35 mg/dL	-	Low
Gestation in Previous Pregnancy	-	≥ 2	High Dia BP	High
No of Pregnancies	-	≥ 3	-	
Unexplained Prenatal Loss	-	Presence	-	
Prediabetes	-	Presence	High BMI	

C. Patient-Specific Interpretations

To complement the global insights, we examine the model's local reasoning with SHAP waterfall plots. A waterfall plot (Fig. 19) decomposes one patient's prediction by starting at the baseline E[f(x)] (the model's expected output over the training background) and then applying each feature's SHAP contribution in sequence until it reaches the patient's final prediction f(x). The x-axis represents the model output scale, so movement to the right increases the prediction and movement to the left decreases it; the y-axis lists the patient's features with their observed values. Bars show the size and direction of each contribution, with red increasing risk and blue decreasing risk.

For this patient, a 24-year-old woman with BMI 23.69 and no sedentary lifestyle, PCOS, prediabetes, or family history (see Table VI), the predicted probability of GDM is f(x) = 0.001, well below the dataset baseline E[f(x)] = 0.612, indicating a low-risk profile. The model arrives at this result through several negative contributions: the absence of unexplained prenatal loss (−0.09), BMI below 25, systolic blood pressure 117 mmHg (below 130 mmHg), and OGTT 140.32 mg/dL (below clinical and data-driven thresholds). Additional decreases come from HDL 53 mg/dL (above threshold) and diastolic blood pressure 80 mmHg (below 85 mmHg).

Her parity of two is below the higher-risk cutoff of three, and the absence of PCOS further reduces risk. Although hemoglobin is 9.8 g/dL, which is below the range where positive SHAP effects were observed in this dataset (>16 g/dL), it does not raise predicted GDM risk here; however, it remains clinically low and may warrant separate attention during pregnancy.

This case is consistent with the data-driven thresholds that we presented earlier in Table V: none of her values exceed the risk cutoffs, and several interacting protective features (such as high HDL, normal blood pressure, and OGTT below threshold) contribute cumulatively to a low-risk prediction. This case highlights how the model synthesizes individual-level clinical factors to make well-reasoned predictions. It also demonstrates SHAP's utility in explaining not only high-risk cases but also in identifying low-risk profiles by tracing the impact of protective factors.

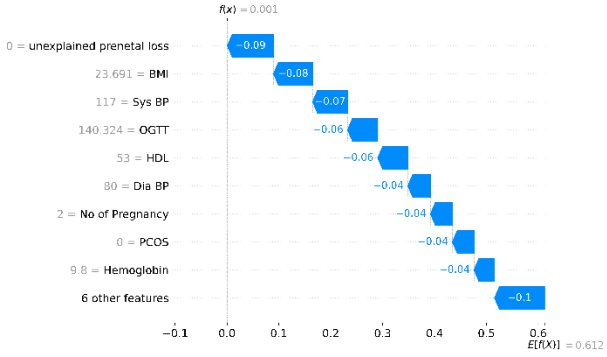


Fig. 19. SHAP waterfall plot for a representative patient.

TABLE VI. PATIENT-LEVEL FEATURE VALUES FOR LOCAL EXPLANATION

Feature	Value
Age	24
No. of Pregnancy	2
Gestation in previous pregnancy	None
BMI	23.96
HDL	53
Family History	None
Unexplained prenatal loss	None
Large Child or Birth Default	None
PCOS	None
Sys BP	117
Dia BP	80
OGTT	140.32
Hemoglobin	9.8
Sedentary Lifestyle	None
Prediabetes	None

VI. Discussion

This work demonstrates how interpretable ensemble modelling, implemented via the EYE-GDM system, can support risk prediction in the context of GDM. We specifically selected GDM as the application focus for several reasons. First, GDM presents a clinically significant yet temporally constrained condition, with substantial risks for both mother and infant, including preeclampsia, macrosomia, neonatal hypoglycaemia, and long-term progression to type 2 diabetes [4]. Its transient onset during pregnancy necessitates timely and explainable risk assessments that clinicians can trust.

From a scientific perspective, the contributions of this work lie in the methodological refinements introduced in EYE-GDM, including the reconfigured ensemble structure, the explanation-weighted meta-learning approach, and the domain-aligned interpretability evaluation. These represent technical extensions of the EYE framework. The applied contribution lies in demonstrating these methodological advances within the maternal health domain, where EYE-GDM provides clinically interpretable risk assessments, data-driven thresholds, and patient-level insights specific to gestational diabetes.

EYE-GDM supports interpretation at both the dataset and patient levels. At the dataset level, SHAP summary plots and dependence plots allow for the identification of global risk patterns, data-driven thresholds, and feature interactions. These insights capture not only the individual impact of clinical features but also how their interactions influence GDM risk. Such results can inform strategic clinical guidelines, such as refining screening protocols for high-risk groups. In contrast, patient-level explanations, visualized through SHAP waterfall plots, help clinicians understand why a specific patient is predicted to be at risk, thereby supporting individualized intervention planning. Together, these interpretation levels enhance the model's clinical utility and support population-level decision-making and personalized care.

The predictive performance of EYE-GDM is competitive, achieving an accuracy of 0.9789, an F1-score of 0.9790, a precision of 0.9747, a recall of 0.9833, and an AUC-ROC of 0.9981. These results outperform previously reported models on the same dataset while simultaneously offering transparent explanations. Compared with conventional ensemble or deep-learning models used for GDM prediction, EYE-GDM provides distinct advantages by integrating interpretability directly within the training process rather than applying it as a separate post-hoc step. This ensures that explanations remain consistent with the ensemble's internal reasoning. In addition, the model combines global and patient-level reasoning within a single framework, enabling both population insights and individual clinical interpretation, which is rarely achieved by previous methods. Furthermore, incorporating expert clinical feedback during validation strengthens the model's credibility and aligns its explanations with real-world diagnostic reasoning and clinical practice.

More importantly, the interpretations derived from SHAP analysis are well-aligned with established medical knowledge, confirming the validity of the model's outputs. SHAP values for Age increase after 40, consistent with the clinical evidence that advanced maternal age is a GDM risk factor [37], [38]. Risk rises significantly for BMI values above 25, aligning with WHO thresholds for overweight [32], [33]. OGTT values show a sharp SHAP increase beyond 160 mg/dL, supporting diagnostic thresholds for impaired glucose tolerance [32]. Hemoglobin contributes to risk in the 13–16 g/dL range only when OGTT is also elevated, underscoring the need to interpret it contextually. HDL levels above 35 mg/dL exhibit protective SHAP values, matching its known cardiometabolic role [36]. Similarly, Sys BP starts contributing to higher risk beyond 130 mmHg, with stronger effects in patients who have low HDL. This pattern reflects the well-known clustering of high blood pressure, abnormal lipids, and elevated glucose seen in metabolic syndrome [39]. Prediabetes and PCOS both contribute positively to GDM risk, consistent with medical literature [40], [41]. Finally, obstetric history factors such as prior gestation with complications and a higher number of pregnancies display threshold-based patterns, reinforcing their importance in GDM prognosis [42], [43].

The interpretability results revealed clinically consistent yet subtle patterns. Interactions such as BMI with HDL and blood pressure with metabolic indicators showed that GDM risk emerges from the combined effects of metabolic and cardiovascular factors rather than single predictors. The model also identified context-dependent relationships, where certain features influence risk only in combination with others, reflecting the complexity of gestational physiology. The close alignment between model-derived thresholds (BMI 25 kg/m², Dia BP 85 mmHg, Sys BP 130 mmHg) and established reference ranges supports the clinical credibility of EYE-GDM outputs. These findings illustrate how interpretable AI can mirror clinical reasoning in maternal health, providing insights that clarify how metabolic, cardiovascular, and obstetric factors interact to shape individual risk. In practice, the identified thresholds can inform more targeted antenatal screening and help refine early intervention strategies for high-risk patients. The variability observed in lifestyle and family-history features highlights the need for broader, standardized datasets to enhance robustness and ensure the generalizability of model insights across populations.

Clinician validation further reinforced these interpretations. The clinicians confirmed that the model's predictions were consistent with established GDM risk factors and reflected the multifactorial nature of clinical practice, where risk factors interact rather than act independently. They noted, however, that hemoglobin alone is not an independent risk factor for GDM, whereas HbA1c is a more reliable predictor. The observed association between elevated hemoglobin levels and higher OGTT values in this dataset was considered important, potentially reflecting physiological mechanisms such as increased blood viscosity [31], suggesting that these factors act in combination rather than through hemoglobin alone. The clinicians also emphasized the need for clearer dataset documentation and recommended involving clinical experts in the data collection process to ensure the inclusion of key GDM biomarkers, such as HbA1c and different OGTT test types. The validation was regarded as valuable both for strengthening trust and for guiding iterative refinement.

While the model showed strong alignment with clinical reasoning, certain limitations remain. This work extends prior

validation of the EYE framework (EYE-WD) to a distinct clinical population, demonstrating its adaptability across domains. However, because the dataset represents a single regional population, sociocultural and healthcare-access factors that influence screening and diagnosis may have affected model behavior. Future work will focus on validating the model across multi-centre and demographically diverse GDM cohorts as such datasets become available.

The results of this work highlight the importance of embedding interpretability directly into model design for maternal healthcare. EYE-GDM demonstrates how interpretable models can uncover clinically relevant patterns that align with established medical knowledge, advancing maternal health AI by combining high predictive accuracy with clinically grounded explanations.

VII. Conclusion and Future Work

EYE-GDM, an implementation of the EYE framework for gestational diabetes risk prediction, delivers strong predictive performance together with clinically meaningful and interpretable outputs. It directly addresses two long-standing gaps in prior work: the tendency to prioritize predictive accuracy without producing explanations that clinicians can apply, and the reliance on post-hoc interpretability methods that fail to capture how models reason internally. By embedding SHAP explanations into the training pipeline, EYE-GDM ensures that each base learner is explained individually and that their contributions are integrated through logistic regression weights, producing ensemble-level reasoning consistent with the model's internal decision logic.

EYE-GDM achieved competitive results, surpassing earlier reports on the same dataset. More importantly, its interpretable outputs were consistent with established medical knowledge, showing that the framework can provide reliable risk prediction while also generating explanations of practical value in maternal care. This combination strengthens confidence in the system as a tool that clinicians can trust.

Beyond its technical contributions, EYE-GDM advances the field of explainable healthcare AI by demonstrating how interpretability can be embedded throughout the model development process rather than appended after training. This integration shifts explainability from a diagnostic tool to a design principle, enabling models that reason transparently while maintaining predictive strength. Within maternal health, EYE-GDM illustrates how such integration can transform complex risk prediction into a clinically meaningful, trust-enhancing process, setting a foundation for future interpretable systems across other sensitive domains.

Future work will continue engagement with clinical experts to refine interpretability outputs and ensure their clarity in practice. Additional directions include extending the framework to other healthcare conditions where interpretability is essential, optimizing it for real-time deployment, and evaluating its performance across diverse populations.

The EYE framework builds on our earlier FIXAIH roadmap, which set out six design priorities for developing trustworthy and interpretable AI in healthcare. While FIXAIH defines high-level goals, EYE operationalizes selected elements into a structured and testable framework centered on interpretability by design, transparent development, and clinical validation. EYE-GDM demonstrates how these principles can be applied in maternal health, while EYE-WD shows their application in women's diabetes. Together, these implementations highlight how domain-specific frameworks can serve as practical steps toward the broader FIXAIH vision, linking conceptual guidance with deployment-ready systems that advance clinical care while maintaining transparency and trust.

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] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," Nat Med, vol. 28, no. 1, pp. 31–38, 2022, doi: 10.1038/s41591-021-01614-0.

[2] P. Kumar, S. Chauhan, and L. K. Awasthi, "Artificial Intelligence in Healthcare: Review, Ethics, Trust Challenges & Future Research Directions," Eng Appl Artif Intell, vol. 120, p. 105894, Apr. 2023, doi: 10.1016/J.ENGAPPAI.2023.105894.

[3] Y. Du, A. M. Antoniadi, C. McNestry, F. M. McAuliffe, and C. Mooney, "The Role of XAI in Advice-Taking from a Clinical Decision Support System: A Comparative User Study of Feature Contribution-Based and Example-Based Explanations," Applied Sciences, vol. 12, no. 20, 2022, doi: 10.3390/app122010323.

[4] World Health Organisation (WHO), "Diabetes," WHO. Accessed: Mar. 01, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes

[5] S. Feduniw, D. Sys, S. Kwiatkowski, and A. Kajdy, "Application of artificial intelligence in screening for adverse perinatal outcomes: A protocol for systematic review," Medicine, vol. 99, no. 50, p. e23681, Dec. 2020, doi: 10.1097/MD.0000000000023681.

[6] K. S. Lee and K. H. Ahn, "Application of artificial intelligence in early diagnosis of spontaneous preterm labor and birth," Diagnostics, vol. 10, no. 9, p. 733, 2020, doi: 10.3390/diagnostics10090733.

[7] M. Becker et al., "Revealing the impact of lifestyle stressors on the risk of adverse pregnancy outcomes with multitask machine learning," Front Pediatr, vol. 10, p. 933266, 2022, doi: 10.3389/fped.2022.933266.

[8] H. Y. Lu et al., "Digital Health and Machine Learning Technologies for Blood Glucose Monitoring and Management of Gestational Diabetes," IEEE Rev Biomed Eng, vol. 17, pp. 98–117, 2024, doi: 10.1109/RBME.2023.3242261.

[9] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," Artif Intell, vol. 267, pp. 1–38, 2019, doi: 10.1016/j.artint.2018.07.007.

[10] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," Lancet Digit Health, vol. 3, no. 11, pp. e745–e750, Nov. 2021, doi: 10.1016/S2589-7500(21)00208-9.

[11] S. Alghamdi, R. Mehmood, F. Alqurashi, T. Alghamdi, A. AlAhmadi, and S. Ghazali, "EYE and EYE-WD: Clinically Validated, Interpretable Ensemble Learning for Women's Diabetes," SSRN Preprint, Aug. 2025, doi: 10.2139/ssrn.5386582.

[12] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use," in Proceedings of the 4th Machine Learning for Healthcare Conference, F. Doshi-Velez, J. Fackler, K. Jung, D. Kale, R. Ranganath, B. Wallace, and J. Wiens, Eds., in Proceedings of Machine Learning Research, vol. 106. PMLR, Sep. 2019, pp. 359–380. [Online]. Available: https://proceedings.mlr.press/v106/tonekaboni19a.html

[13] A. Sumathi, S. Meganathan, and B. V. Ravisankar, "An Intelligent Gestational Diabetes Diagnosis Model Using Deep Stacked Autoencoder.," Computers, Materials & Continua, vol. 69, no. 3, pp. 3109–3126, Mar. 2021, doi: 10.32604/cmc.2021.017612.

[14] A. Sumathi and S. Meganathan, "Ensemble Classifier Technique to Predict Gestational Diabetes Mellitus (GDM).," Computer Systems Science Engineering, vol. 40, no. 1, pp. 313–325, 2022, doi: 10.32604/csse.2022.017484.

[15] R. Jader and S. Aminifar, "Predictive Model for Diagnosis of Gestational Diabetes in the Kurdistan Region by a Combination of Clustering and Classification Algorithms: An Ensemble Approach," Applied Computational Intelligence and Soft Computing, vol. 2022, no. 1, 2022, doi: 10.1155/2022/9749579.

[16] H. Gallardo-Rincón et al., "MIDO GDM: an innovative artificial intelligence-based prediction model for the development of gestational diabetes in Mexican women," Sci Rep, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-34126-7.

[17] T. Zheng et al., "A simple model to predict risk of gestational diabetes mellitus from 8 to 20 weeks of gestation in Chinese women," BMC Pregnancy Childbirth, vol. 19, no. 1, 2019, doi: 10.1186/s12884-019-2374-8.

[18] J. Shen et al., "An innovative artificial intelligence-based app for the diagnosis of gestational diabetes mellitus (GDM-AI): Development study," J Med Internet Res, vol. 22, no. 9, p. e21573, Sep. 2020, doi: 10.2196/21573.

[19] Y. T. Wu et al., "Early Prediction of Gestational Diabetes Mellitus in the Chinese Population via Advanced Machine Learning," Journal of Clinical Endocrinology & Metabolism, vol. 106, no. 3, 2021, doi: 10.1210/clinem/dgaa899.

[20] X. Hu, X. Hu, Y. Yu, and J. Wang, "Prediction model for gestational diabetes mellitus using the XG Boost machine learning algorithm," Front Endocrinol (Lausanne), vol. 14, 2023, doi: 10.3389/fendo.2023.1105062.

[21] M. Watanabe, A. Eguchi, K. Sakurai, M. Yamamoto, and C. Mori, "Prediction of gestational diabetes mellitus using machine learning from birth cohort data of the Japan Environment and Children's Study," Sci Rep, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-44313-1.

[22] Y. Du, A. R. Rafferty, F. M. McAuliffe, L. Wei, and C. Mooney, "An explainable machine learning-based clinical decision support system for prediction of gestational diabetes mellitus," Sci Rep, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-05112-2.

[23] V. Vivek Khanna et al., "Explainable artificial intelligence-driven gestational diabetes mellitus prediction using clinical and laboratory markers," Cogent Eng, vol. 11, no. 1, p. 2330266, Dec. 2024, doi: 10.1080/23311916.2024.2330266.

[24] H. Zaky et al., "Machine learning based model for the early detection of Gestational Diabetes Mellitus," BMC Med Inform Decis Mak, vol. 25, no. 1, p. 130, 2025, doi: 10.1186/s12911-025-02947-3.

[25] S. Alghamdi, R. Mehmood, F. A. Alqurashi, and A. Alzahrani, "Paving the Roadmap for XAI and IML in Healthcare: Data-Driven Discoveries and the FIXAIH Framework," IEEE Access, vol. 13, pp. 174393–174427, 2025, doi: 10.1109/ACCESS.2025.3616353.

[26] A.sumathi and S.Meganathan, "Gestational Diabetes Mellitus (GDM Dataset)." Accessed: Jul. 10, 2025. [Online]. Available: https://www.kaggle.com/dsv/3245285

[27] S. Sperandei, "Understanding logistic regression analysis," Biochem Med (Zagreb), vol. 24, no. 1, pp. 12–18, 2014, doi: 10.11613/BM.2014.003.

[28] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proceedings of the 31st International Conference on Neural Information Processing Systems, in NIPS'17. , Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 4768–4777.

[29] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nat Mach Intell, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/S42256-019-0138-9.

[30] S. H. Golden, C. Yajnik, S. Phatak, R. L. Hanson, and W. C. Knowler, "Racial/ethnic differences in the burden of type 2 diabetes over the life course: a focus on the USA and India," Diabetologia, vol. 62, no. 10, pp. 1751–1760, 2019, doi: 10.1007/s00125-019-4968-0.

[31] S. Rafaqat and S. Rafaqat, "Role of hematological parameters in pathogenesis of diabetes mellitus: A review of the literature," World Journal Hematology, vol. 10, no. 3, pp. 25–41, Mar. 2023, doi: 10.5315/WJH.V10.I3.25.

[32] American Diabetes Association Professional Practice Committee, "2. Diagnosis and Classification of Diabetes: Standards of Care in Diabetes—2025," Diabetes Care, vol. 48, no. Supplement_1, pp. S27–S49, Jan. 2025, doi: 10.2337/dc25-S002.

[33] World Health Organization, "Obesity and overweight," May 2025. Accessed: Sep. 18, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

[34] National Board of Medical Examiners, "Laboratory Reference Values," Mar. 2025. Accessed: Oct. 18, 2025. [Online]. Available: https://www.nbme.org/laboratory-values

[35] P. K. Whelton et al., "2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines," J Am Coll Cardiol, vol. 71, no. 19, pp. 2199–2269, May 2018, doi: 10.1016/j.jacc.2017.11.005.

[36] Cleveland Clinic, "HDL: Why It's 'Good' Cholesterol." Accessed: Sep. 10, 2025. [Online]. Available: https://my.clevelandclinic.org/health/articles/24395-hdl-cholesterol

[37] A. Bouzaglou et al., "Pregnancy at 40 years Old and Above: Obstetrical, Fetal, and Neonatal Outcomes. Is Age an Independent Risk Factor for Those Complications?," Front Med (Lausanne), vol. 7, p. 208, May 2020, doi: 10.3389/fmed.2020.00208.

[38] E. A. AlJahdali and N. S. AlSinani, "Pregnancy outcomes at advanced maternal age in a tertiary Hospital, Jeddah, Saudi Arabia," Saudi Med J, vol. 43, no. 5, p. 491, May 2022, doi: 10.15537/smj.2022.43.5.20220023.

[39] J. Kim, "Metabotype Risk Clustering Based on Metabolic Disease Biomarkers and Its Association with Metabolic Syndrome in Korean Adults: Findings from the 2016–2023 Korea National Health and Nutrition Examination Survey (KNHANES)," Diseases, vol. 13, no. 8, p. 239, 2025, doi: 10.3390/diseases13080239.

[40] G. Wilkie, E. Delpapa, and H. Leftwich, "Early Diagnosis of Prediabetes among Pregnant Women that Develop Gestational Diabetes Mellitus and Its Influence on Perinatal Outcomes," Am J Perinatol, vol. 41, no. 3, pp. 343–348, 2024, doi: 10.1055/A-1682-2643.

[41] H. J. Kim, E. H. Kim, E. Ko, S. Park, and Y. Lee, "The Impact of Polycystic Ovary Syndrome on Gestational Diabetes Mellitus, Disease Knowledge, and Health Behaviors," Healthcare, vol. 13, no. 7, 2025, doi: 10.3390/healthcare13070717.

[42] J. Pukkila et al., "The recurrence risk of gestational diabetes according to the number of abnormal values in the oral glucose tolerance test," Acta Obstet Gynecol Scand, vol. 104, no. 8, pp. 1452–1462, Aug. 2025, doi: 10.1111/AOGS.15148.

[43] B. Liu et al., "Higher Numbers of Pregnancies Associated With an Increased Prevalence of Gestational Diabetes Mellitus: Results From the Healthy Baby Cohort Study," J Epidemiol, vol. 30, no. 5, pp. 208–212, 2020, doi: 10.2188/jea.JE20180245.

An Intelligent Platform for Behavior Modification and Office Syndrome Risk Reduction Using MediaPipe and Computer Vision

Sumran Chaikhamwang[1], Wijitra Montri[2], Chalida Janthajirakowit [3], Srinuan fongmanee[4]
Information Technology-Faculty of Digital Technology, Chiangrai Rajabhat University, Thailand[1, 2, 3]
Computer Science-Faculty of Digital Technology, Chiangrai Rajabhat University Dept., Thailand[4]

***Abstract*—Office Syndrome, a musculoskeletal disorder prevalent among office workers, poses significant risks to health, productivity, and quality of life. Traditional preventive approaches, such as ergonomic guidelines and reminder-based systems, often fail due to limited user adherence and practicality. To address this gap, this study developed an intelligent platform that integrates MediaPipe and computer vision to monitor sitting posture, eye-to-screen distance, and sitting duration in real time. The system provides automated notifications and stretching recommendations, combining detection, feedback, and behavioral intervention into a sensor-free and cost-effective solution. The platform was evaluated in terms of technical performance and user behavioral impact. Results demonstrated high system accuracy, with the eye-distance detection module achieving 95.2% accuracy, followed by long sitting alerts (92.5%) and proximity alerts (90.1%). User evaluations confirmed that real-time notifications increased awareness and encouraged healthier working behaviors. These findings highlight the potential of computer vision based approaches for ergonomic health promotion. The proposed platform contributes not only to preventive strategies for Office Syndrome but also to advancing user-centered, technology-driven health solutions adaptable to both office and remote work environments. This study not only demonstrates technical performance but also introduces a novel integration of Media Pipe based posture and facial detection with behavioral modification features, which previous ergonomic systems have not addressed. The proposed framework contributes a new perspective for integrating real time feedback with computer vision in promoting sustainable ergonomic behavior.**

Keywords*—*Office syndrome; computer vision; MediaPipe; behavior modification, ergonomics

I. Introduction

The rapid transformation of global economies and technological advancements in the 21st century has significantly reshaped working patterns, particularly among office workers who spend extended hours in front of computer screens. One of the major health consequences of this lifestyle is Office Syndrome, a musculoskeletal disorder resulting from prolonged static postures, repetitive movements, and the overuse of specific muscle groups. This condition not only causes discomfort but has also emerged as a public health issue that adversely affects work performance, quality of life, and healthcare costs [1], [2].

Recent studies have shown that over 70% of office workers worldwide report musculoskeletal discomfort, particularly in the neck, back, and shoulders, and these numbers are continuously increasing due to the widespread adoption of remote and hybrid working models [3], [4]. In Thailand, the Department of Disease Control reported that more than 60% of the working-age population suffers from Office Syndrome, especially individuals between 30–44 years old, with a rising annual trend of approximately 10% [5]. These statistics highlight the urgency of preventive solutions that are both accessible and adaptable to modern working environments.

Although ergonomic guidelines and workplace exercise programs have been promoted for decades, evidence indicates that long-term adherence remains limited. For instance, interventions that rely solely on user self-discipline—such as scheduled stretching breaks or posture adjustments—tend to fail when workloads are heavy or deadlines are pressing [6]. In response, researchers have developed mobile and desktop applications to support healthy work behaviors. Many of these tools use simple timers or reminder notifications to encourage breaks, while others employ wearable devices to track sitting duration and activity [7], [8]. However, such approaches remain insufficient. Reminder-based systems are often ignored, and wearable devices require additional hardware that may not be practical or affordable for all users [9].

In recent years, computer vision based methods have gained attention as they allow automated posture detection without additional sensors. Applications using pose estimation frameworks such as OpenPose or MediaPipe have demonstrated high accuracy in detecting improper postures [10]. Nevertheless, most existing studies have focused on posture recognition alone, without integrating real-time behavioral interventions or personalized exercise recommendations [11], [12]. For example, Zhang et al. (2024) combined MediaPipe with YOLOv5 to monitor joint motion in patients, but the system lacked behavior modification features [13]. Similarly, ergonomic assessment tools using MediaPipe Pose provided valuable risk analysis but stopped short of offering proactive, user-centered feedback for long-term behavior change [14].

This gap reveals a crucial research opportunity: while current technologies can detect risky behaviors, they do not effectively translate detection into sustainable prevention. To address this challenge, the present study proposes an intelligent

platform that integrates MediaPipe based posture and facial detection with real-time notification systems and guided stretching recommendations. Unlike conventional reminder applications, this approach combines detection, alerting, and intervention into a comprehensive solution that is sensor-free, cost-effective, and adaptable to both office and home-working contexts. By filling this gap, the system not only advances technological applications for health promotion but also offers a practical contribution to reducing healthcare costs and improving worker productivity in the long term.

This study contributes to bridging the gap between detection and actual behavioral change by developing an intelligent, vision based system that not only monitors ergonomic risks but also delivers personalized, real-time interventions. Unlike prior MediaPipe-based systems that focused solely on posture detection, this platform emphasizes a behavioral modification mechanism and human-centered feedback, enabling sustainable improvements in ergonomic habits.

II. Related Works

A. Definition and Causes of Office Syndromeelecting

Office syndrome refers to a cluster of musculoskeletal symptoms associated with work, primarily resulting from repetitive postures, prolonged improper sitting, and inadequate physical activity [6], [7]. The main causes are prolonged computer use in inappropriate postures, such as neck bending, staring too closely at screens, or placing the wrists at improper angles, which lead to overuse of specific muscle groups and inflammation [8], [9]. Additionally, psychological factors, such as work-related stress, can exacerbate pain symptoms [10]. If left unaddressed, these conditions can negatively affect work efficiency and overall quality of life [11].

This syndrome is not limited to office workers but also affects those working from home, who often lack ergonomically suitable equipment, significantly increasing the risk of musculoskeletal pain [12]. Understanding the causes and mechanisms of office syndrome is therefore crucial for developing effective prevention and treatment strategies. Current research emphasizes analyzing work behaviors and environmental factors to identify solutions tailored to individual contexts [13].

B. Incidence and Prevalence of Office Syndrome

The incidence of office syndrome has been steadily increasing worldwide, posing a significant occupational health challenge. International studies report that over 71.9% of office employees experience work-related musculoskeletal disorders (WMSDs), particularly lower back and neck pain [12]. In India, 63% of regular computer users reported musculoskeletal discomfort [14]. The COVID-19 pandemic and the shift to remote work have further increased the prevalence due to a lack of ergonomically sound work environments [8].

In Thailand, data from the Department of Disease Control indicate that over 60% of the working-age population is affected, especially those aged 30–44, a demographic with high work intensity [9], [15]. The annual increase in cases, averaging 10%, highlights the urgent need for preventive measures [15]. The widespread prevalence of office syndrome has broader socioeconomic implications, including reduced productivity and increased absenteeism [11].

C. Impacts of Office Syndrome on Health and Work Efficiency

Office syndrome affects not only physical health but also mental well-being and work performance [11]. Chronic pain can lead to fatigue, lack of concentration, and irritability, directly impacting work quality, increasing errors, and prolonging task completion time [10]. Moreover, pain from office syndrome negatively affects sleep quality, leading to insufficient rest, which further exacerbates symptoms [7]. These effects illustrate a chain reaction in which physical discomfort contributes to psychological stress, ultimately impairing work performance [12].

Long-term consequences can include severe conditions such as herniated discs or tendon inflammation, requiring costly and time-consuming treatments [13]. Organizations are also affected through reduced employee productivity and increased sick leave, resulting in economic losses at both micro and macro levels [8]. Prevention is therefore a more cost-effective approach than treatment, emphasizing the need for tools that support daily preventive practices [14].

D. Strategies for Preventing and Addressing Office Syndrome

Globally recognized strategies for preventing office syndrome include modifying the work environment according to ergonomic principles and adjusting personal behaviors [9], [7]. Properly positioning chairs, desks, and computer screens reduces musculoskeletal strain [13], [8]. Additionally, exercise programs that promote stretching and core strengthening are important for minimizing injury risk [6]. However, maintaining consistent adherence to these practices is challenging for most workers [11].

In the digital era, technology plays a vital role in addressing these issues. Applications or software that remind users to take breaks or adjust postures have gained attention [16]. Several studies indicate that applications providing continuous guidance and notifications significantly reduce musculoskeletal pain and improve work behaviors [14], [10], [12], aligning with the lifestyle needs of modern workers [12].

E. Technology Applications for Preventing Office Syndrome

Technology has become a key tool in managing and preventing office syndrome effectively. Developing smartphone or computer applications using AI or sensors to analyze work postures and provide real-time feedback is increasingly popular in current research [6]. Such applications alert users when they maintain improper postures for extended periods and suggest exercise videos tailored to individual needs [16], facilitating automated and continuous prevention [8], [13].

In addition, wearable devices, such as smartwatches that monitor movement and heart rate, offer promising options for detecting risky behaviors and notifying users [14]. Integrating these technologies with ergonomic principles can create effective tools to promote worker health [10]. Further research

and development in this area remain crucial to meet the diverse needs of users [11].

F. Application of MediaPipe and Computer Vision in Preventing Office Syndrome

The integration of advanced technologies, such as MediaPipe and computer vision, has shown great promise in mitigating the risks associated with Office Syndrome. MediaPipe, a framework developed by Google, enables real-time pose estimation and gesture recognition, facilitating the development of applications that monitor and correct users' postures during prolonged computer use. For example, Zhang et al. demonstrated the effectiveness of combining MediaPipe with YOLOv5 for assessing the range of motion in patients with spinal diseases and frozen shoulder, highlighting the potential of such technologies for musculoskeletal health monitoring [17].

In addition, computer vision techniques have been applied to automate ergonomic risk assessments. Singhtaun et al. employed MediaPipe Pose to develop an automated Rapid Entire Body Assessment (REBA) system, which identifies and mitigates ergonomic risks in real-time [18]. These applications are particularly valuable in workplace settings, where continuous monitoring and immediate feedback are crucial to prevent musculoskeletal disorders.

Wearable devices, such as smartwatches, have also been explored to monitor users' physical activity and posture. These devices provide real-time alerts and feedback, encouraging users to adopt healthier habits and reducing the risk of developing Office Syndrome. Research indicates that integrating wearable technology with ergonomic interventions can significantly enhance users' physical well-being and work performance [19].

Furthermore, recent studies have explored the use of computer vision and MediaPipe in various real-time monitoring scenarios, including detecting postural deviations, estimating joint angles, and providing corrective feedback for seated office workers [20], [21]. The integration of these technologies into workplace health strategies provides a proactive approach for Office Syndrome prevention, promoting a healthier work environment and enhancing employee productivity [22].

III. Method

This study followed a Research and Development (R&D) approach with the goal of creating and testing an intelligent platform that helps reduce the risk of Office Syndrome. The process was carried out in four stages, described below.

A. Data Collection and Review

In this stage, the researchers conduct a comprehensive review to gain a deep understanding of both the problem and the relevant technologies. The study is divided into two main parts: theoretical and technical reviews.

The theoretical review focuses on literature related to Office Syndrome, including its causes, symptoms, prevalence, and impacts on health and work performance [23], [24]. Insights from this review serve as the foundation for determining the necessary features and functions of the application.

The technical review investigates the technologies and tools available for development, particularly image-processing libraries and techniques that can detect users' posture and behavior. Popular frameworks such as MediaPipe and OpenCV are explored for their capabilities in analyzing human physical characteristics [25], [26]. This review ensures the selection of the most suitable and effective tools for application development.

B. Application Design

Based on the review, we designed a system with four main functions:

- Detecting sitting posture and facial orientation.
- Measuring the distance between the eyes and the screen.
- Sending real-time notifications when risky behavior is detected.
- Suggesting stretching exercises to relieve strain.

The user interface (UI) was designed to be simple and friendly, so office workers could easily integrate it into daily routines. A database was also designed to store posture data, screen distance records, and activity logs for later analysis [28]-[29].

C. Application Development

The system was built in modular components:

- Posture Detection Module – used MediaPipe Pose (BlazePose model), which identifies 33 body landmarks, to monitor sitting posture. We also applied MediaPipe Face Mesh to estimate facial orientation and screen distance through eye tracking. We set the detection confidence thresholds at 0.5 for both detection and tracking, after testing different values.
- Notification Module – generated alerts when users sat too long or were too close to the screen. Users could adjust the alert frequency, typically every 30–60 minutes.
- Data Management Module – stored all posture, distance, and sitting duration logs in a MySQL database for later evaluation.

Testing setup, we used a Logitech HD 1080p webcam (1920×1080 pixels, 30 fps) in an office-like room with normal lighting. Each test session lasted about 45 minutes, designed to resemble typical office work conditions.

D. Testing and Evaluation

We evaluated the platform in two ways: technical performance and behavioral impact.

1) Technical performance

- We compared the system's detection results with manual coding by human observers.
- Three standard evaluation metrics were used [30].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

where,

TP = True Positives

TN = True Negatives

FP = False Positives

FN = False Negatives.

To ensure generalizability, the evaluation included 30 participants with different physical builds and working behaviors.

2) Behavioral evaluation

- We assessed how the system influenced users' habits through the System Usability Scale (SUS) [27], and a short Likert-scale questionnaire adapted from ergonomic behavior assessments.
- Questions asked about perceived usefulness, ease of use, and effectiveness in promoting healthier behaviors.
- We also conducted short interviews, giving participants space to share how they felt about the notifications, exercise suggestions, and overall usability.

This mixed evaluation allowed us to measure not only how accurate the system was technically, but also how well it worked in encouraging actual behavioral change. The steps for face detection, eye distance detection, and sitting position detection can be explained in the flowchart shown in Fig. 1 to 3, respectively.

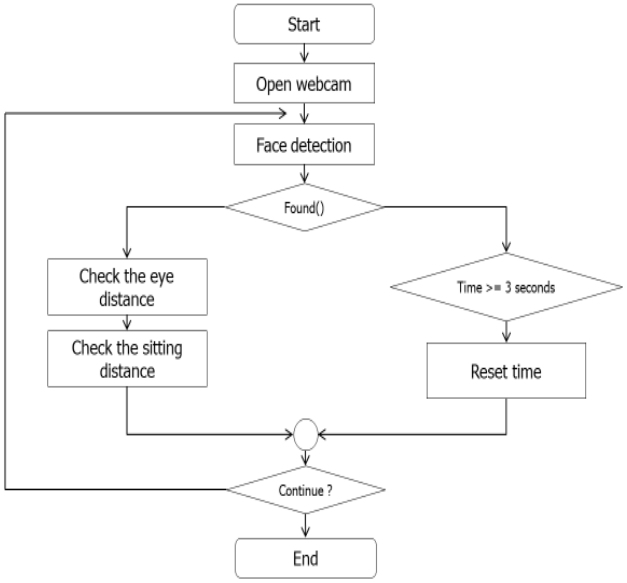


Fig. 1. Face verification flowchart.

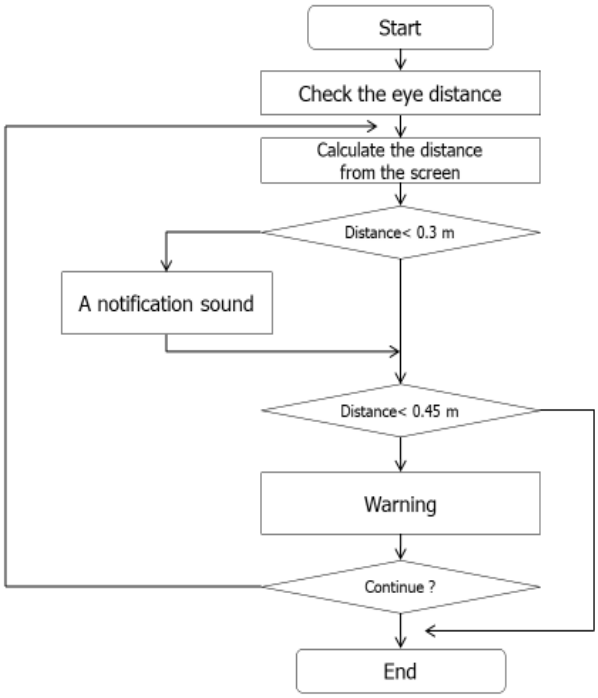


Fig. 2. Flowchart of visual distance inspection.

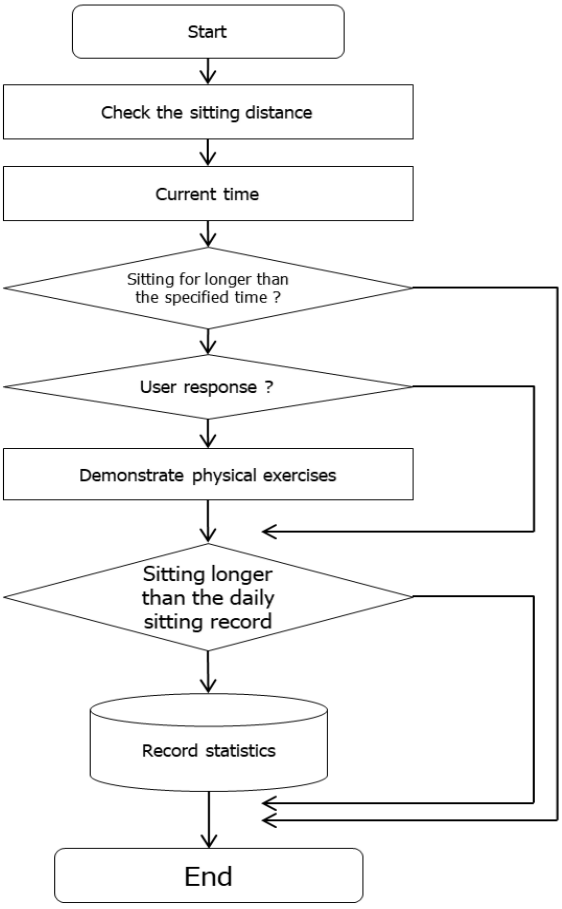


Fig. 3. Flowchart of sitting inspection.

E. Novelty and System Contribution

The novelty of this research lies in the integration of MediaPipe's real-time pose estimation with a dual-layer behavioral intervention model. The system uniquely combines detection, notification, and guided exercise recommendation without relying on external sensors. Moreover, the inclusion of user feedback analytics enables the platform to act as a continuous behavior modification assistant rather than a one-time monitoring tool, setting it apart from previous ergonomic studies.

IV. Results

The research findings are presented in two main components: the User Interface (User) and the System Administration (Backend). Each component was designed to accommodate specific functions and usability requirements, as detailed below.

A. User Interface (User)

The user interface was designed with a total of ten primary screens, allowing users to conveniently access all core functionalities of the application. The home screen acts as the central hub, displaying real-time face detection, eye-to-screen distance, and both the current continuous sitting duration and the longest sitting duration recorded for the day. Additionally, the home screen provides menu options and control buttons to activate or deactivate the webcam, thereby giving users flexibility in managing system usage.

The interface consists of the following screens: Home, Notification Settings, Proximity Alerts, Sitting Duration Alerts, Website Homepage, Muscle Exercise Demonstration, Additional Exercise Demonstrations, Exercise Detail View, Article List, and Article Detail View. These screens collectively ensure that users receive comprehensive support, ranging from ergonomic monitoring to knowledge enhancement. Illustrative examples of these screens are provided in Fig. 4 to 8.

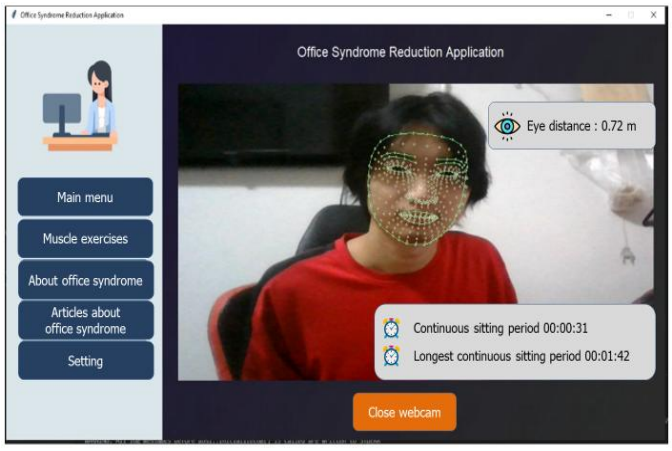


Fig. 4. Application main page.

As shown in Fig. 4, the main page contains the following components menu, face detection display, real-time eye distance, current continuous sitting duration, longest continuous sitting duration today, and webcam on/off button.

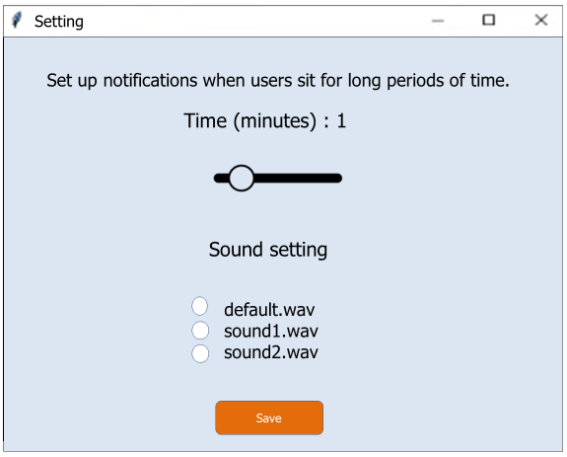


Fig. 5. Sound setting.

As shown in Fig. 5, the notification settings interface, which is designed to allow users to customize the notification system according to their preferences. The main components are as follows:

- Sitting duration alert timer: Users can conFig. the notification interval within a range of 30–60 minutes. This feature helps reduce the health risks associated with prolonged sitting.
- Notification sound options: The system provides multiple alert sounds, enabling users to select the most suitable tone based on their preferences and convenience.
- Save settings button: Users can save their customized settings so that the system retains and applies them in subsequent notifications.
- The notification settings interface thus functions as an essential tool that empowers users to manage their work habits more flexibly and effectively. By adjusting both the frequency and the style of notifications, individuals can align the system with their specific work patterns and personal needs.

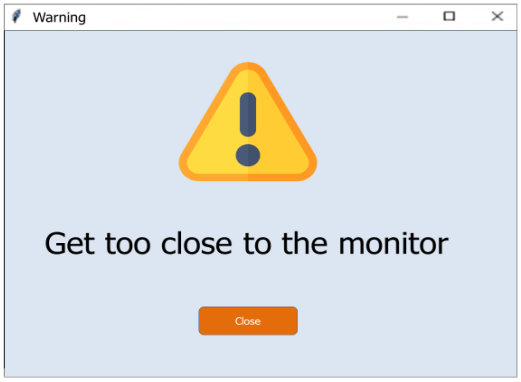


Fig. 6. Warning for users approaching the screen too close.

As shown in Fig. 6, if the user approaches the screen too close (30-45 centimeters), the system will alert the user to adjust their distance from the screen.

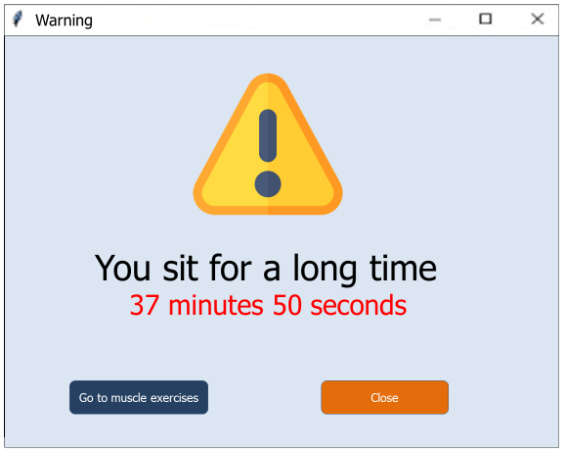


Fig. 7. Long sitting alert.

Fig. 7 shows the long sitting alert page. When a user sits for longer than the set time, an alert will appear with two buttons: "Go to Exercise Page" and "Close." If the user selects "Go to Exercise Page," the system will automatically open the exercise website, as shown in Fig. 8 shows web application section.

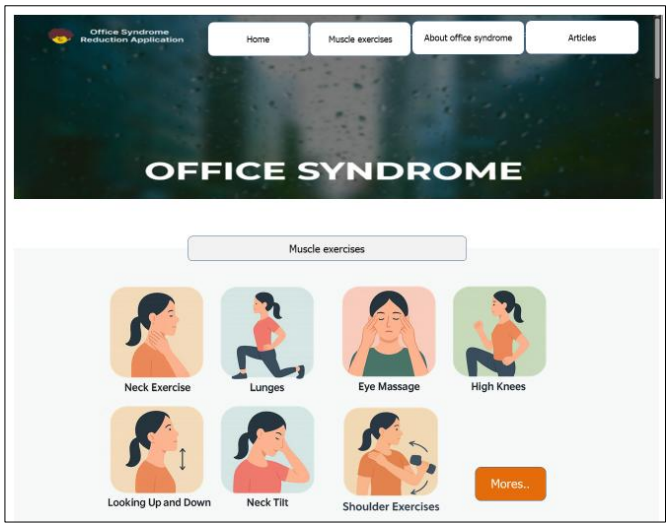


Fig. 8. Web application section.

Fig. 8 presents user recommendations for various muscle-building activities. The exercises listed in the database include examples of Neck Exercise, Lunges, Eye Massage, High Knees, Looking Up and Down, Neck Tilt, and Shoulder Exercises. When users click on a specific exercise, the exercise details are displayed, including text, images, audio, and video.

B. System Administration (Backend)

The system administration (backend) was developed to facilitate administrative operations and ensure secure as well as efficient management of application data. It is composed of three primary components:

1) Login page: Administrators are required to provide a valid username and password in order to access the system. The authentication process is carried out to verify the accuracy of the credentials, thereby maintaining data security. Once access is granted, a navigation menu is displayed on the left-hand side of the interface, offering options for managing both articles and exercise content.

2) Data management page: Within this interface, records such as articles and exercise routines are presented in a structured list format. Functional controls are provided, enabling administrators to add, delete, or update information. This design ensures that the process of content management is both straightforward and effective.

3) Add/edit data page: This page allows administrators to create new records or modify existing ones. The interface supports the uploading of image files and the input of detailed descriptions with formatting options. Upon submission, the system automatically updates the database and manages the associated image files to maintain consistency and accuracy of stored data.

An example of the system administration (backend) interface is illustrated in Fig. 9.

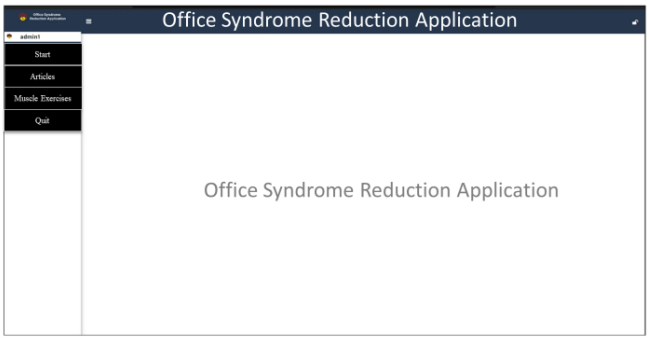


Fig. 9. Administrator page.

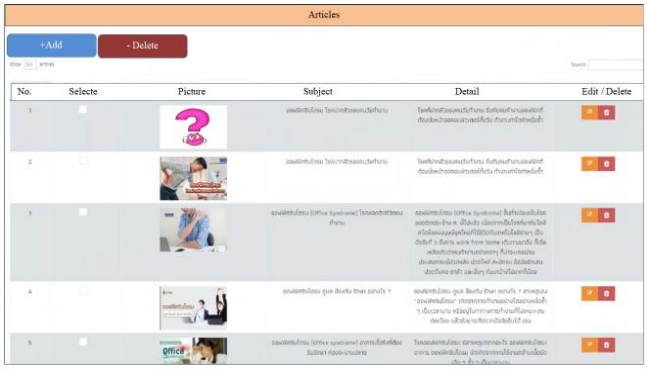


Fig. 10. Data management page.

As shown in Fig. 10, the list page contains the name of the currently opened item and the detailed content displayed. As in the example image, it is an article with content, images, and buttons for adding, deleting, and editing.

As shown in Fig. 11, you can enter information to add to the system. An image file can be uploaded to the host and various formatting options are available. Once you've completed the information, scroll down to the bottom of the page and you'll see the "Save Data" button.

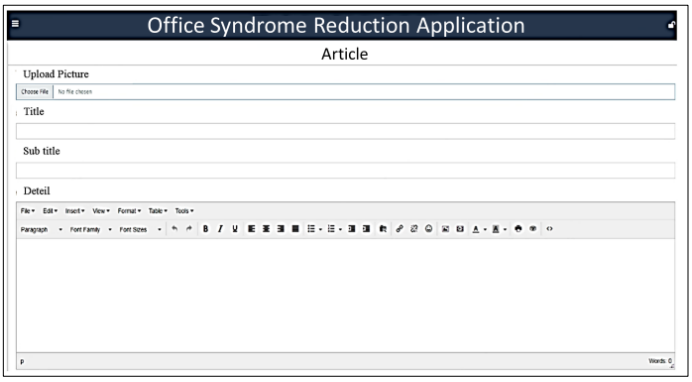


Fig. 11. Data addition page.

TABLE I. SYSTEM PERFORMANCE

Module/Feature	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Eye distance	95.2	94.8	96.1	95.4
Long sitting alert	92.5	91.7	93	92.3
Approaching the screen too close	90.1	89.5	90.8	90.1

The system performance was evaluated using Accuracy, Precision, Recall, and F1-Score across three key modules: Eye Distance, Long Sitting Alert, and Approaching the Screen Too Close, as summarized in Table I and the related graphs.

- The Eye Distance module demonstrated the highest performance (Accuracy 95.2%, Precision 94.8%, Recall 96.1%, F1-Score 95.4%), indicating that the system can accurately and reliably measure the distance between the eyes and the screen.
- The Long Sitting Alert module achieved slightly lower but still robust performance (Accuracy 92.5%, Precision 91.7%, Recall 93%, F1-Score 92.3%), showing that prolonged sitting can be effectively detected and notified.
- The Approaching the Screen Too Close module recorded the lowest scores among the three (Accuracy 90.1%, Precision 89.5%, Recall 90.8%, F1-Score 90.1%), yet its performance remains sufficient for practical use.

Overall, the system provides a balanced and reliable detection of all three events, offering both high accuracy and comprehensive coverage, making it well suited for real world applications. As shown in Fig. 12.

To validate the proposed system, its accuracy was benchmarked against existing ergonomic monitoring models such as OpenPose-based detection [17] and sensor-based wearable systems [19]. The results indicate comparable or superior accuracy (average 92.6%) while eliminating the need for external devices, confirming the efficiency and practicality of the camera-only solution.

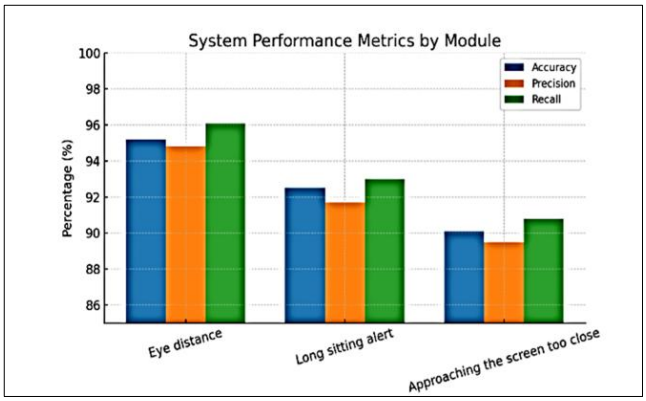


Fig. 12. System performance metrics by module.

V. DISCUSSION

The findings of this study demonstrate that the integration of MediaPipe and computer vision techniques provides an effective means for monitoring sitting posture, eye-to-screen distance, and sitting duration in real time. The system's ability to generate timely notifications and recommend stretching exercises highlights its potential as a preventive tool against Office Syndrome.

One important observation is that users reported increased awareness of their posture and a willingness to adjust their behavior when notifications were triggered. This result is consistent with previous studies, which emphasized the role of real-time feedback in promoting ergonomic practices in the workplace [23], [27]. Moreover, the modular design of the system, which separates detection, notification, and data management, allows for scalability and future integration with emerging technologies such as wearable devices or cloud-based analytics.

Despite its promising results, the system has certain limitations. Environmental conditions such as poor lighting or low-resolution cameras can reduce detection accuracy. In addition, while the system provides basic stretching recommendations, it does not yet personalize these suggestions according to user-specific needs or health conditions. Future research should explore adaptive algorithms that can tailor interventions based on user profiles and longitudinal behavior tracking.

Overall, this study confirms that computer vision based ergonomic systems can play a significant role in workplace health promotion. By addressing the identified limitations and incorporating advanced personalization, the system could be further enhanced for large-scale deployment in organizational and home-office settings.

The observed improvements in user posture awareness align with findings by [8], who emphasized that real-time feedback significantly reduces musculoskeletal discomfort. Unlike timer-based systems, the proposed platform dynamically adapts notifications to user posture, leading to sustained behavioral change. This alignment with prior ergonomic theories supports the platform's potential as a preventive health technology.

A. Related Works Comparison

To highlight the contributions of this research, a comparison was made between existing studies and the proposed system. Table II presents a summary of the key features of related works.

TABLE II. COMPARISON OF RELATED WORKS

Author/Year	Technology Used	Features Provided	Limitations	Contribution of this Study
Sarma et al. (2022) [25]	MediaPipe, OpenCV	Posture detection in real time	No alert system, no behavior modification	Added notification and exercise guidance
Koul et al. (2021) [26]	Computer Vision (Pose Estimation)	Ergonomic risk assessment	Limited to static evaluation	Integrated real-time monitoring and alerts
Hassan et al. (2024) [23]	Ergonomic Survey + CV	Identified risk factors in office workers	Based on self-report, not automated	Automated detection with objective data
Okunola et al. (2020)	Wearable Devices	Tracked sitting duration and activity	Required external devices	Camera-based, no wearables needed
Proposed Study (2025)	MediaPipe + OpenCV + Notifications	Real-time face/posture detection, alerts, stretching guidance, data management	Sensitive to lighting and camera quality	Comprehensive, modular, user-friendly system

VI. CONCLUSION

This research demonstrates the effectiveness of integrating MediaPipe and computer vision techniques into an intelligent platform for preventing Office Syndrome. The system's ability to detect posture, monitor eye to screen distance, and identify prolonged sitting combined with real-time notifications and stretching recommendations proved to be both accurate and practical. Technical evaluation indicated robust performance across all modules, while user feedback confirmed increased awareness of ergonomics and positive behavior modification. Importantly, the system operates without the need for wearable devices, making it an accessible and cost-effective solution for diverse working contexts.

Nevertheless, certain limitations remain, particularly in sensitivity to lighting conditions and the absence of personalized exercise recommendations. Future work should focus on adaptive algorithms, integration with wearable devices, and cloud-based analytics to enhance personalization and scalability. Overall, the study underscores the potential of computer vision based systems to transform workplace ergonomics by fostering sustainable behavioral change, reducing health risks, and improving overall productivity.

In summary, the study not only provides a validated prototype for ergonomic risk detection but also contributes to the broader field of human computer interaction by integrating behavioral modification strategies into computer vision systems. Future studies should explore AI driven personalization and large scale longitudinal trials to assess long-term behavioral outcomes and workplace health impact.

REFERENCES

[1] O. O. Chinedu, A. T. Henry, J. J. Nene, and J. D. Okwudili, "Work-related musculoskeletal disorders among office workers in higher education institutions: A cross-sectional study," Ethiopian Journal of Health Sciences, vol. 30, no. 5, 2020, doi: 10.4314/ejhs.v30i5.10.

[2] N. Narainsamy, F. F. Akpa-Inyang, S. C. Onwubu, N. Govender, and J. D. Pillay, "Ergonomic challenges and musculoskeletal pain during remote working: A study of academic staff at a selected university in South Africa during the COVID-19 pandemic," Int. J. Environ. Res. Public Health, vol. 22, no. 1, p. 79, 2025, doi: 10.3390/ijerph22010079.

[3] J. W. Kim, J. Y. Choi, E. J. Ha, and J. H. Choi, "Human pose estimation using MediaPipe pose and optimization method based on a humanoid model," Applied Sciences, vol. 13, no. 4, p. 2700, 2023, doi: 10.3390/app13042700.

[4] J. W. Kim, J. Y. Choi, E. J. Ha, and J. H. Choi, "Human pose estimation using MediaPipe pose and optimization method based on a humanoid model," Applied Sciences, vol. 13, no. 4, p. 2700, 2023, doi: 10.3390/app13042700.

[5] B. Demissie, G. B. Mekonnen, and T. M. Aytenew, "Musculoskeletal disorders among computer user workers in Ethiopia: A systematic review and meta-analysis," Scientific Reports, vol. 15, no. 1, p. 26858, 2025, doi: 10.1038/s41598-025-96470-0.

[6] Department of Medical Services, Ministry of Public Health, Thailand, Guidelines for Preventing Work-Related Musculoskeletal Disorders. Bangkok, Thailand, 2025.

[7] C. M. Lind, F. Abtahi, and M. Forsman, "Wearable motion capture devices for the prevention of work-related musculoskeletal disorders in ergonomics—An overview of current applications, challenges, and future opportunities," Sensors, vol. 23, no. 9, p. 4259, 2023, doi: 10.3390/s23094259.

[8] J. Rabal-Pelay, C. Cimarras-Otal, S. Coscolla-Palma, A. Burillo-Villuendas, J. A. Villalba-Ruete, and A. V. Bataller-Cervero, "Effect of active breaks on stress and musculoskeletal discomfort during work in office workers," Sport Sciences for Health, pp. 1–9, 2025, doi: 10.1007/s11332-025-01488-6.

[9] I. N. Gomez, C. G. Suarez, K. E. Sosa, and M. L. Tapang, "Work from home-related musculoskeletal pain during the COVID-19 pandemic: A rapid review," Int. J. Osteopath. Med., vol. 47, p. 100654, 2023, doi: 10.1016/j.ijosm.2022.12.001.

[10] I. W. Suryasa, M. Rodríguez-Gámez, and T. Koldoris, "The COVID-19 pandemic," Int. J. Health Sci., vol. 5, no. 2, p. 572194, 2021, doi: 10.53730/ijhs.v5n2.2937.

[11] R. Leyshon, K. Chalova, L. Gerson, A. Savtchenko, R. Zakrzewski, A. Howie, and L. Shaw, "Ergonomic interventions for office workers with musculoskeletal disorders: A systematic review," Work, vol. 35, no. 3, pp. 335–348, 2010, doi: 10.3233/WOR-2010-0994.

[12] B. Demissie, E. T. Bayih, and A. A. Demmelash, "A systematic review of work-related musculoskeletal disorders and risk factors among computer users," Heliyon, vol. 10, no. 3, 2024, doi: 10.1016/j.heliyon.2024.e25075.

[13] W. Zhang et al., "Combined MediaPipe and YOLOv5 range of motion assessment system for spinal diseases and frozen shoulder," Scientific Reports, vol. 14, no. 1, p. 15879, 2024, doi: 10.1038/s41598-024-66221-8.

[14] C. Singhtaun, S. Natsupakpong, and P. Lorprasertkul, "Ergonomic risk assessment using human pose estimation with MediaPipe pose," in Proc. 2024 7th Artificial Intelligence and Cloud Computing Conf., Dec. 2024, pp. 465–471, doi: 10.1145/3719384.3719453.

[15] C. H. Huang, L. E. Yen, and Y. T. Chen, "Development of an ergonomic risk assessment system for repetitive tasks by skeleton recognition based on MediaPipe," in IET Conf. Proc. CP929, vol. 2025, no. 15, pp. 137–139, Apr. 2025, doi: 10.1049/icp.2025.2511.

[16] S. Sultana, M. S. H. Hriday, R. Haque, and M. Das, "Health monitoring through wearables: A systematic review of innovations in cardiovascular

disease detection and prevention," Strategic Data Management & Innovation, vol. 2, no. 1, 2025, doi: 10.71292/sdmi.v2i01.13.

[17] T. Agostinelli, A. Generosi, S. Ceccacci, and M. Mengoni, "Validation of computer vision based ergonomic risk assessment tools for real manufacturing environments," Scientific Reports, vol. 14, no. 1, p. 27785, 2024, doi: 10.1038/s41598-024-79373-4.

[18] V. Wagh, M. W. Scott, and S. N. Kraeutner, "Quantifying similarities between MediaPipe and a known standard to address issues in tracking 2D upper limb trajectories: Proof of concept study," JMIR Form. Res., vol. 8, no. 1, p. e56682, 2024, doi: 10.2196/56682.

[19] F. Kahal, A. Al Darra, and A. Torbey, "Computer vision syndrome: A comprehensive literature review," Future Sci. OA, vol. 11, no. 1, p. 2476923, 2025, doi: 10.1080/20565623.2025.2476923.

[20] A. Carrasco-Uribarren et al., "A computer vision based application for the assessment of head posture: A validation and reliability study," Applied Sciences, vol. 13, no. 6, p. 3910, 2023, doi: 10.3390/app13063910.

[21] K. Totlani et al., "Real-time human pose estimation using MediaPipe and artificial intelligence applications in health and fitness," in Proc. 2024 4th Int. Conf. Artif. Intell. Signal Process. (AISP), Oct. 2024, pp. 1–6, doi: 10.1109/AISP61711.2024.10870725.

[22] B. Cardoso, C. Mateus, R. Magalhães, and M. A. Rodrigues, "Ergonomic intervention program for office workers: A case study about its effect in computer vision syndrome and musculoskeletal discomfort," Ergonomics, vol. 68, no. 1, pp. 51–62, 2025, doi: 10.1080/00140139.2023.2288543.

[23] R. Noy, Y. Shkedy, and Y. Vaisbuch, "Ergonomic injuries and their complications: Prevention and treatment," Oper. Tech. Otolaryngol. Head Neck Surg., 2025, doi: 10.1016/j.otot.2025.05.006.

[24] J. Bort-Roig et al., "A digital intervention to reduce home-office workers' sedentary behaviour: Protocol for the evaluation of the Click2Move programme, a cluster randomised controlled trial," BMC Public Health, vol. 25, no. 1, p. 387, 2025, doi: 10.1186/s12889-025-21598-7.

[25] N. Abulail, A. Y. Owda, and M. Owda, "Real-time detection and analysis of facial movement disorders using MediaPipe and OpenCV: A Streamlit-based framework," in Proc. 2025 12th Int. Conf. Inf. Technol. (ICIT), May 2025, pp. 13–18, doi: 10.1109/ICIT64950.2025.11049234.

[26] U. U. Deshpande et al., "A review of machine learning techniques for ergonomic risk assessment based on human pose estimation," Preprint, 2025, doi: 10.21203/rs.3.rs-5654648/v1.

[27] J. J. M. Ecleo et al., "Evaluation of the usability and user experience of a digital platform for mental health assessment," Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 3, 2025, doi: 10.14569/ijacsa.2025.0160310.

[28] J. Zeng et al., "A real-time posture monitoring system based on Kalman-MediaPipe fusion algorithm," in Proc. 2025 5th Int. Symp. Comput. Technol. Inf. Sci. (ISCTIS), May 2025, pp. 923–930, doi: 10.1109/ISCTIS65944.2025.11065400.

[29] S. Beddar-Wiesing, A. Moallemy-Oureh, M. Kempkes, and J. M. Thomas, "Absolute evaluation measures for machine learning: A survey," arXiv preprint arXiv:2507.03392, 2025, doi: 10.48550/arXiv.2507.03392.

[30] M. Chabbouh, S. Bechikh, E. Mezura-Montes, and L. Ben Said, "Evolutionary optimization of the area under precision-recall curve for classifying imbalanced multi-class data," J. Heuristics, vol. 31, no. 1, p. 9, 2025, doi: 10.1007/s10732-024-09544-z.

A Review of Ransomware Detection Models for Cybersecurity Driven IIoT in Cloud Environments

Abrar Ali, Norah Hamed, Monir Abdullah
College of Computer Science and Information Technology, University of Bisha, Bisha 61922, Saudi Arabia

***Abstract*—Ransomware is currently one of the most severe cybersecurity threats and not only attacks legacy systems but cloud systems and Industrial Internet of Things (IIoT) systems as well. Security and privacy threats are heightened as these systems integrate more closely and thus are exposed to sophisticated and long-lasting attacks. This paper provides a comprehensive review of ransomware prevention and detection measures in cloud and IIoT environments with an emphasis on the usage of Machine Learning (ML) and Deep Learning (DL) models. Research studies published across IEEE, Elsevier, and Springer databases between 2020 and 2024 were analyzed. Our check reveals Ensemble methods and Random Forest (RF) are two of the ML methods most in use, with each at 18.00%, followed by Neural Networks (NNs) at 12.00%, with older models such as Support Vector Machines (SVMs) with 10.00%, Naïve Bayes (NBs) had 7.00%, and Decision Trees (DTs) still in use with utilization at 9.00% . Additionally, DL approaches (including Convolutional NN (NN), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), and Recurrent NN (RNN)) account for 20.00% of the techniques deployed, highlighting their growing prominence in IIoT security and ransomware research. Indicative of their integration into hybrid ML pipelines, Light Gradient Boosting Machine (LightGBM) and other ensemble boosting frameworks comprise 16.00%. Last but not least, other novel and specialized models including Extreme Gradient Boosting (XGBoos), Self-Organizing Maps (SOM), Gain Ratio, and Digital DNA account for 8.00% of the overall utilization observed throughout study. Among DL methods, Recurrent NNs (RNNs) are at the forefront with 40%, followed by CNNs with 30%, CNN–RNN hybrid models at 20%, and Autoencoders with 10%. Integration of cryptographic schemes, federated learning, blockchain-based audit mechanisms, and adaptive runtime mechanisms have further boosted the mechanisms of anomaly detection with detection rates of over 99% for polymorphic and zero-day ransomware.**

***Keywords*—Ransomware; Industrial Internet of Things; cloud computing; machine learning; deep learning; blockchain**

I. Introduction

It is any program engineered to disturb the usual operation of an operating system, literally through penetrating or exploiting holes in pre-existing software in a bid to impair a computer's operation through its resource, network, and data. Among the most common malware typologies are viruses, worms, spyware, Trojan horses, adware, and ransomware. Today, ransomware is also deemed the most prevalent malware for having a characteristic called cryptovirology, where it uses encryption methods to invade user data and computer files. Some of the latest significant ransomware attacks include those of Taiwan Semiconductor Company from WannaCry, SamSam Ransomware, WannaCry against the U.K. National Health Service, and the ransomware attack on Foxconn. The United States is the country most targeted in cybercrime, where government sector organizations are the highest victims, then the manufacturing, education, and healthcare industries [1].

A. Cloud Computing

Cloud computing offers a low-cost and scalable storage and data processing solution for vast amounts of data, and edge computing offers computing power and data storage where they are needed locally [2]. Artificial Intelligence (AI) and ML [3] enhance these systems.

This research offers a holistic evaluation of cryptographic methods applied in cloud computing to overcome data security issues. Among the most notable approaches reviewed is DNA cryptography, popular for high storage density and high security, though currently in a nascent developmental state; Elliptic Curve Cryptography (ECC), popular for security through a small key size, hence suitable for resource-limited spaces such as the IoT; and homomorphic encryption, allowing computations on encrypted data without decrypting it, hence maintaining data confidentiality in processing. Additionally, hybrid cryptography combines symmetric and asymmetric encryption for increased efficiency and security, while lightweight cryptography is specifically designed for use in devices with limited computational power, such as sensors and Radio Frequency Identification (RFID) devices. Comparative studies of these approaches have been carried out, basing analyses on performance, security aspects, and usability, with recommendations to merge methods such as ECC and homomorphic encryption with blockchain technology for increased robustness. Reliability and accuracy for these processes are determined based on application, with success rates of 85% to 98%, depending on applied algorithms [4].

B. Vulnerabilities and Technology Used

This study examines program protection for Autonomous Vehicular Cloud Computing (AVCC) platforms against timing side-channel attacks. A compiler-level obfuscation preprocessor is introduced by the researchers, adeptly rewriting the input program's control flow dynamically and in a randomized manner. It takes advantage of the Low Level Virtual Machine (LLVM) compiler for executing conditional branch transformations and extraneous code insertion for an increase in logical complexity and reverse engineering attempts deterrents. Independently of input language and platform, the system is centered on ARM-based embedded platforms and extends branch conversion opportunities. Experimental evaluations

validate higher runtime variability in various program versions, thereby enhancing protection against the timing analysis attacks. Having low overhead, the software solution is a quick and efficient method for protecting AVCC platforms [5].

Defines data storage and sharing security in cloud methods such as cryptography, access control, ML differential privacy, watermarking, and probabilistic methods. The essay compares techniques in detail, strength and weakness. Privacy and security are achieved by cryptography but are susceptible to exposure after encryption key hacking. Access control minimizes data leakages without any transformation cost but fails to detect malicious agents. Differential privacy is computationally expensive but preserves data utility and privacy. Probabilistic methods and watermarking are efficient in tracing leakers, and watermarking is good for tracing culprits and probabilistic methods are resistant to data alteration. The paper highlights the point that no single solution offers comprehensive security, in favor of composite solutions for offering comprehensive protection. Results show significant data security improvement, with performance relying on the approach and use [6].

Aims to provide a decentralized, privacy-preserving public auditing protocol that allows cloud data integrity while taking potential auditor or cloud server manipulation into account. The system employs the utilization of blockchain technology to provide secure, random challenges and audit results logging in a transparent manner to enable tamper-proof operations. Zero-knowledge proofs are utilized for the protection of user data privacy during auditing, while the Proof of Work mechanism prohibits the cloud server from preparing challenges in advance. With a decentralized system, the system avoids relying on fully trusted third parties. The outcomes validate the effectiveness of the system in data privacy protection and public, traceable auditing. Security analysis assures its resilience against guess attacks on challenge messages, and experimental tests on the ethereum test net confirm its practicability with high efficiency and minimal communication and computational overhead [7].

A blockchain-supported certificate-less public cloud data integrity audit scheme, proposed specifically for safe cloud storage while eliminating issues such as troublesome certificate administration and key escrow. It requires the administration of semi-trusted Third-Party Auditors (TPAs) through blockchain technology and preservation of user data privacy in the course of audit. A novel data structure in combination with a counting Bloom filter and Multi-Merkel hash tree is developed to enable efficient dynamic update of data. System robustness for audit correctness, privacy preservation, and preservation against substitution attack is experimentally tested. Experimental verification proves its efficiency with 99% detection likelihood for data tampering and low computational and communication overheads, rendering it plausible and practical [8].

The project develops Run-time Adaptations for Data protection (RADAR), a data protection system for dynamic cloud settings that adapts to dynamically emerging threats and configurations in real-time. Key methods utilized are Models@Runtime for runtime modeling of the system, pattern matching for identifying malicious configurations, adaptation rules for autonomous reconfigurations, and search algorithms (e.g., Best-First Search) for optimization of adaptation. Further functionalities such as Topology and Orchestration Specification for Cloud Applications (TOSCA) for system topology modeling, Eclipse Modeling Framework (EMF) for models at runtime, and encryption for data protection are included. The results establish the efficacy of RADAR in identifying and rectifying data protection threats such as General Data Protection Regulations (GDPR) compliance in low cost and low loss of functionality. High scalability and accuracy of the system were experimentally confirmed through numerous case studies, rendering the system applicable in complex real-world situations [9].

The study intends to deploy an intelligent behavior-based malware detection system in a cloud computing environment to detect known and unknown malware with high accuracy. Suspicious files are uploaded to the cloud, stored in VMs, and their unique behaviors and traits are tracked. The system employs modern methods, including rule-based detection, ML classifiers (J48, RF and K-Nearest Neighbors (KNN)), and the Cloud-Based Behavior Model (CBCM), to identify whether or not files contain malware. Beyond the state of the art, the evaluation reveals that low false positive rates (0.4% for ML and 6.6% for rule-based detection) and high detection rates (up to 99.8%) are achieved. The system is very accurate (99.7%) and scalable and can therefore be deployed in dynamic and complex environments [10].

This study focuses on advanced malware detection methodologies in cloud infrastructure. It aims at minimizing the threats of advanced malware, such as polymorphic and metamorphic variants, from evading simplistic signature-based detection schemes. It integrates ML algorithms, DL models, and behavioral detection for differentiating malicious from benign code. Principal methods involve both dynamic and static analyses for feature extraction, virtual machine introspection (VMI), and heuristic detection schemes. Test cases establish high detection rates with low false positives and thereby system scalability and performance in detecting known entities and unknown threats. It offers a comprehensive solution with high security for complex cloud infrastructure [11].

LSTM and BILSTM models are being studied for real-time malware detection in clouds using RNNs. To find malicious activity in VMs, it employs system monitoring of Central Processing Unit (CPU), memory, and disk utilization. A data set of 113 unique malware specimens collected in a real cloud infrastructure environment yielded 40,680 test instances. Both models achieved performance values above 99% in accuracy, precision, recall, and F1-score, with LSTMs requiring less training time. The results demonstrate how RNN-based methods can be used to identify sophisticated malware threats on cloud devices [12]. Table I summarizes the vulnerabilities and technologies used for prevention the ransomware in cloud computing.

TABLE I. VULNERABILITIES AND TECHNOLOGY USED FOR RANSOMWARE PREVENTION IN CLOUD COMPUTING

Ref	Vulnerabilities in Cloud Computing	Technology Used in Prevention	Year
[4]	Data security issues (data exposure, unauthorized access)	DNA Cryptography, ECC, homomorphic encryption, hybrid and lightweight cryptography	2024
[5]	Timing side-channel attacks in Autonomous Vehicular Cloud Computing	LLVM-based compiler obfuscation (branch transformation, junk code)	2022
[6]	Exposure of encryption keys, undetected malicious agents, data leakage	Cryptography, Access control, ML-based differential privacy, watermarking, probabilistic methods	2022
[7]	Manipulation by cloud servers/auditors, lack of trust in third parties	Blockchain-based decentralized auditing, Zero-knowledge proofs, Proof of Work	2020
[8]	Inconvenient certificate management, key escrow issues	Blockchain, Certificate-less auditing, Bloom filter, Multi-Merkel hash tree	2023
[9]	Real-time adaptation to emerging threats, dynamic configuration flaws	Models@Runtime, Pattern matching, Adaptation rules, TOSCA, EMF, encryption	2021
[10]	Detection of unknown and polymorphic malware	Behavior-based detection, CBCM model, ML classifiers (J48, RF, and KNN)	2021
[11]	Evasion by smart malware (polymorphic/metamorphic), low detection accuracy	ML/DL models (CNN and RNN), Dynamic & static analysis, Virtual Machine Introspection, Heuristic analysis	2024
[12]	Malware behaviors in cloud VMs, performance monitoring	RNNs (LSTM and BiLSTM) analyzing CPU, memory, disk usage	2021

II. RELATED WORKS

A. Industrial Internet of Things (IIoT)

IIoT is a fusion of IoT technology and industrial devices applied in industry [13]. presents a new cryptographic algorithm using the application of the Harris Hawks Optimization (HHO) approach, combined with feature selection, to enhance communication efficiency and security for communication between IIoT devices and cloud computing platforms [14]. This sector has attracted much attention in recent years with its ability to transform industry manufacturing using intelligent and efficient methods. The IIoT transformed industry settings through the linking of a wide range of devices in a network and facilitating the collection, analysis, and decision-making processes in real-time. Data-informed decisions have seen increased business efficiency in operation, decreased downtime, and enhanced productivity [15]. Traditional IIoT architecture, based mainly on cloud computing, however, is faced with inherent shortcomings in addressing the high volume and velocity of IIoT data, bandwidth limitations, and confidentiality-related data issues [16]. To overcome these issues, distributed edge-to-cloud computing has been introduced as a viable solution through the blending of edge computing and cloud computing in supporting IIoT endeavors.

B. Vulnerabilities and Technology Used

The title introduces a novel ransomware detection framework customized for IIoT networks. Principal Contributions Revolutionary FL Design: An asynchronous peer-to-peer FL structure does not entail the involvement of a cloud server. Enhanced Data Processing: CDAE obtains adversarial robustness against IID and non-IID data. Scalability and Privacy: Lowers bandwidth consumption and maintains data privacy throughout training and update processes. We tested the model on three data sets (X-IIoTID, ISOT, and NSL-KDD), and it showed the following features: Enhanced accuracy, recall, and F1 values compared to current ransomware detection models, including those for unknown variants (evasion attacks).Efficient working on IID and non-IID data distributions, maintaining adaptability in real-world heterogeneous settings. Adversarial robustness against attacks such as Fast Gradient Sign Method (FGSM) and brute-force (BF) attacks [17]. key issue of the protection of IIoT devices from every kind of cyberattack and introduces an AI-motivated professional system for detection. Key Contributions Classification of IIoT Attacks: Labels attacks as denial-of-service (DoS), data tampering, device hijacking, and physical tampering. Proposed Expert System: Combines rule-based reasoning, anomaly detection, and reinforcement learning. Utilize characteristics such as "duplication and retransmission rate" for detecting attacks more efficiently. Attack Types Addressed Man-in-the-Middle (MitM), Distributed DoS (DDoS), and Start-Stop attacks. Testing and Verification: Examined on real Programmable Logic Controllers (PLCs) and IIoT protocols such as Modbus, MQTT. Showed high precision (99.7%) and low latency in detecting attacks [18].

This research examines the convergence of ransomware attacks and IoT technologies and provides an in-depth examination of the evolution, classifications, and consequences of ransomware, especially in IoT settings. It categorizes ransomware into two kinds: Crypto-ransomware, where user data is encrypted and a ransom must be paid in exchange for a decryption key, and Locker ransomware, where the user's device is locked until a ransom is paid. It suggests not paying ransom as a way to discourage cybercriminals and promotes cybersecurity best practice, such as disabling of macros, privilege limitation for users, and advanced detection system implementation [19].

Provides a lightweight and efficient means of securing IoT applications via a publish-subscribe communication paradigm. Its primary issues it addresses are the limitations of resource-poor IoT devices, the lack of increased scalability, and the ineffectiveness in resource-poor devices of traditional security schemes such as Transport Layer Security (TLS). Proposed Solution: Architecture: Fog nodes assist in storage and computing. A broker facilitates communication and authentication in the publish-subscribe infrastructure. Encryption: Secret messages are maintained via Advanced Encryption Standard CCM (AES-CCM), and communication expense are reduced. Authentication: A safe key exchange utilizing ECC is utilized and is resource-light compared to

standard public-key schemes. Optimization: Comparatively, handshakes, message lengths, and memory consumption are minimized in this scheme with respect to TLS-dependent schemes. Paper Contributions: It demonstrated a resource-light encryption scheme tailored for resource-limited IoT devices. It ensured end-to-end security via significantly fewer communication, storage, and processing expenses [20].

Model Building: CNN models (1D, 2D, 3D) were developed with input layers, several convolutional layers, dropout layers (for overfitting prevention), and fully connected dense layers. Transfer learning was adopted to enhance classification efficiency for binary and multiclass classification. Paper Datasets: Four available datasets (BoT-IoT, MQTT-IoT-IDS2020, IoT-23, IoT Network Intrusion) were transformed to develop additional datasets, IoT-DS-1 and IoT-DS-2, with increased attack diversity. Characteristics were derived using tools such as CICF low meter, with particular consideration of flow-based traffic characteristics. The developed models performed well on each of the datasets. For instance, the CNN1D model performed up to 99.9% accuracy in certain instances. CNN1D and CNN2D were more accurate and efficient than CNN3D.paper Practical Utility: The generalizability of the developed model suggests potential use in real-world IoT network security situations [21].

Recommends up-to-date DL models for detecting anomalies in IoT networks with particular interest in RNNs models: RNN models use LSTM, BiLSTM, and Gated Recurrent Unit (GRU) architectures and layer normalization, dropout, and regularization to forestall overfitting. Hybrid CNN-RNN models use the CNN layers for extracting spatial features before inputting data into the RNN layers. The scientific paper resulting from the work had the following consequences: High detection rates on various datasets, CNN-BiLSTM having the highest overall detection rates for all cases. Good at identifying novel kinds of attacks through utilizing methods for reducing class imbalances. Lightweight binary classifiers showed promising evidence for real-time anomaly detection for IoT [22]. Considers utilizing Hierarchical Federated Learning (HFL) and Federated Learning (FL) for enhancing Intrusion Detection Systems (IDS) for IoT networks. NSL-KDD was the utilized dataset that contained imbalanced data for five normal and attack sample classes. 122 input features, two hidden layers, and output layer NNs were utilized. Findings of the experiment in the scientific paper were as follows: HFL outperformed FL at every point in accuracy, convergence rate, and non-iid data handling. HFL's hierarchical structure minimized communication overhead and optimized data inconsistency handling. Sting Accuracy: HFL had a 3-6% higher accuracy than FL in a majority of instances [23].

Recent studies have overwhelmingly explored the intersection of IoT and cyber threats, specifically ransomware. Existing literature suggests that the ubiquitous implementation of IoT devices has brought forward newfound vulnerabilities, making devices susceptible to exploitation via ransomware attacks. Scholarly research carried out by Nkenyereye et al. (2022) exhaustively studies the evolution of ransomware, describing the shift in tactics to exploit the inherent properties of IoT ecosystems, such as limited processing power, device heterogeneity, and lack of comprehensive security. The article divides ransomware targeting IoT devices into various kinds, such as crypto-ransomware and locker-ransomware, and studies popular infection vectors, including phishing, malicious codes, and remote exploitation. The article also covers the weaknesses of traditional mitigatory approaches—such as antivirus products and signature-based detection schemes—in adequately confronting threats specific to IoT devices. Valuable contributions of the present study include a review of prevailing mitigatory frameworks and a suggested set of future directions for improvement. Such suggested directions include detection schemes via AI, blockchain technology for communication via secured channels, and the enforcement of device-level tailored security policies specific to IoT devices. Such a study places stress on the need for adaptive and customized security products in response to the prevailing scenario of IoT ransomware [24]. Table II summarizes the vulnerabilities and technologies used for ransomware prevention in IIoT.

TABLE II. VULNERABILITIES AND TECHNOLOGY USED FOR RANSOMWARE PREVENTION IN IIOT

Ref	Vulnerabilities in IIoT	Technology Used in Prevention	Year
[3]	Scalability, security, and legacy system integration	Distributed Edge-to-Cloud computing platforms (e.g., ThingsBoard, Azure IoT)	2024
[13]	IoT ransomware attacks, file encryption, unauthorized modifications	L-IDS (TEE, decoy files, entropy, fuzzy hashing, GNB classifier)	2024
[14]	Secure and efficient communication between IIoT and cloud	Harris Hawks Optimization-based cryptographic algorithm	2024
[15]	Digital forensics issues due to device heterogeneity and decentralization	Blockchain with fuzzy hashing and smart contracts	2021
[16]	Latency, data consumption, and privacy issues	Edge computing (Device, Edge, and Cloud layers)	2020
[17]	Centralized detection model vulnerabilities, adversarial threats	Asynchronous Peer-to-Peer Federated Learning with CDAE	2021
[18]	DoS, data tampering, hijacking, and physical tampering in IIoT	AI-driven expert system (rule-based, anomaly detection, reinforcement learning)	2024
[20]	IoT resource limitations, inefficiency of TLS	Lightweight encryption (AES-CCM), ECC, fog nodes	2020
[21]	Anomaly detection challenges in diverse IoT networks	CNN-based DL models with flow-feature extraction	2021
[22]	Detection of novel attack types and class imbalance	Hybrid CNN-RNN models with LSTM, BiLSTM, GRU	2022
[23]	Non-IID data handling, communication overhead in IDS	HFL model	2021
[24]	Ransomware threats in IoT due to weak security and phishing	AI detection models, blockchain, device-specific security policies	2021

III. Ransomware Attack

Represent a novel and serious danger to international cybersecurity. Ransomware encrypts data or prevents users from accessing their own systems with the aim of extorting ransom payments, which have considerable financial and functional damage to individuals, businesses, and even governments. With the evolution of ransomware strategies, conventional detection techniques are unable to detect and remove such threats effectively and on time [25]. Ransomware not only targets traditional systems but also IoT systems. These systems are usually made up of many resource-constrained IoT devices with diverse requirements. In addition, in contrast to the majority of conventional computing systems, IoT devices mainly provide services and functionality, as opposed to holding sensitive information [13]. Fig. 1 show the ransomware attack process.

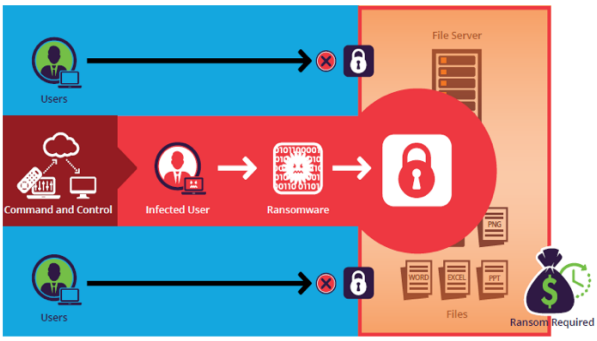


Fig. 1. Ransomware attacks process.

Ransomware is a type of malicious program that can render a computer or data attached to a computer inaccessible or encrypted. It can in some cases completely destroy the system, resulting in the loss of valuable and irretrievable data. Through such a measure of control, cybercriminals demand a heavy ransom from the victim in exchange for regaining access.

1) Crypto-ransomware: describes the amalgamation of ransomware attacks and IoT technology, with a holistic breakdown of the past, classifications, and consequences of ransomware, particularly in the scenario involving IoT.

This paper divides ransomware into two kinds: Crypto-ransomware: It encrypts user files and requests ransom for a decrypt key [19].

2) Locker ransomware: Locks out the user's device, disallowing usage until ransom is paid in return. It advises against ransom payments in a bid to discourage cybercriminals and encourages stringent cybersecurity practice, such as disabling of macros, restricting user permissions, and adopting advanced detection methods [19].

3) Doxware: Threatens to publish the victim's data unless a ransom is paid. The research categorizes ransomware into crypto-ransomware, locker ransomware, doxware, and mobile-based ransomware according to their unique attack patterns and impacts. The research also follows the evolution of ransomware from Crypto Locker (2013) to Ryuk (2020), showing how the tactics of ransomware have developed with encryption and anonymous payment methods. Detection and prevention are carried out through varied methodologies [26].

4) Mobile-based ransomware: This strategy focuses on mobile devices in particular, often through the use of screen locks or data encryption. Such methods based on ML have shown promise in identifying ransomware patterns, up to 97.3% accurate in identification rates by the use of RF, SVM, and NNs. Honey potting methods are also available for passive monitoring through luring ransomware into controlled sections for the sake of analysis. Statistical anomaly detection methods have also been utilized; they are, however, limited in countering more sophisticated, stealthy forms of ransomware [26]. Fig. 2. is an example of Ransomware attacks.

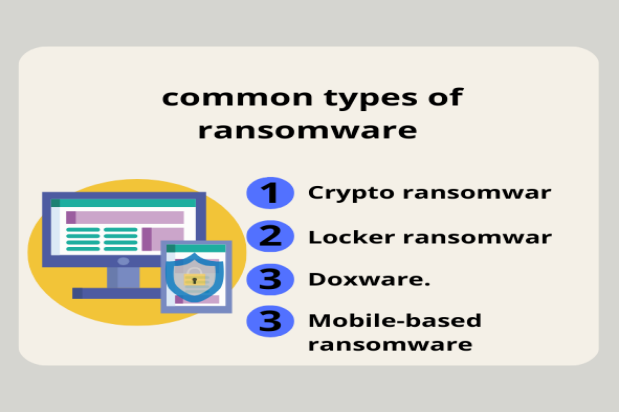


Fig. 2. Example of ransomware attacks.

IV. Ransomware Detection Models

The objective of this research is 1) exploration of the ransomware life cycle on the Windows operating system, 2) behavioral analysis of ransomware specimens in order to draw out various attributes of malicious code schemes, and 3) constructing and validating ML models of ransomware detection on diversified ransomware and non-ransomware specimens [1]. A stacked ensemble learning strategy utilizing six classifiers (Gaussian NB, KNN, DT, Logistic Regression (LR), Multi-Layer Perceptron, and SGD Classifier). A subsequent layer utilizing LightGBM to forecast the identified ransomware into related families [25]. A set of ML models comprising Self-Organizing Maps (SOM), RF classifier, and LSTM networks are utilized for behavior analysis. Implementation of LSTM networks enhances sequential behavior analysis, a key aspect for countering nascent ransomware approaches [26]. Brinkley et al. (2024) have suggested a ML-driven IDS for identifying zero-day ransomware attacks on the basis of as-yet unidentified data. A wide range of varied ML algorithms such as RF, SVM, and NNs were applied in the research and contrasted against a set of 2,850 samples taken from open-source cybersecurity databases and simulation labs. A comprehensive feature engineering mechanism was proposed in the study, with particular focus on system anomalies such as file access behavior, encryption activity, and network behavior. Key findings indicated the NN model had the highest accuracy (92.4%), and RF balanced a blend of accuracy and computational proficiency. A rigorous series of simulations in compliance with zero-day simulations

were carried out on the system and were found to have high generalizability towards hitherto unknown versions of ransomware. In spite of the growth of NN with increased resources and latency, the research proved the feasibility and usability of ML for real-time ransomware detection. It offers a strong framework for adaptive and dynamic cybersecurity protection, as a corrective measure against the inability of signature-based technologies to identify the newer ransomware threats [27]. This study uses XGBoost classifier and RF algorithms for detection and classification of ransomware attacks [28]. We reviewed ransomware's new prevention and detection breakthroughs and suggested directions and challenges for future studies. We actually examined a few well-known ransomware instances and also created our own experimental ransomware, AESthetic, to evade eight well-known antivirus products [29]. Static analysis for ransomware detection. Removal of disassembling procedure through direct feature extraction from raw byte using frequent pattern mining. The Gain Ratio method was implemented to use it on feature selection. RF classifier with in-depth investigation to the impact of both trees and seed amounts in the present study was implemented [30], [31]. Tried the classification algorithm's performance for NBs and RF. The algorithms' performance was assessed on the basis of Accuracy, Precision, Recall, and F-Measure [32]. We presented a comprehensive detail of malware analysis. Cyber Threat Hunting (CTH) methods are narrated on the basis of data analysis technique implemented. Ransomware building and research directions are introduced. Ransomware datasets implemented in the prior studies are cited along with their data sources [33]. AI-driven methods, such as ML and DL, to augment ransomware detection. Conventional ML Models: SVM, RF, NBs, DTs. DL models: CNNs, BiLSTMs, Auto-encoders. Ensemble Learning: A gathering of various classifiers for boosting detection rates. Real-Time Detection: AI models having runtime detection capabilities for ransomware [34]. Examined are the models:

- Blockchain-Based Data Backup and Recovery – Effective for safe, immutable backups but expensive to retain.
- Smart Contract-Enabled Ransomware Detection – Provides for automatic detection and response but is not scalable as it demands high computational requirements.
- Blockchain-Enhanced Secure Communication Channels – Blocks ransomware propagation through encrypted messaging but is not scalable as it faces scalability issues.
- Decentralized Malware Analysis Platforms – Effective for threat intelligence exchange but requires high computational infrastructure and facilities [35]. A combination of several methods, including behavioral analysis, heuristic detection, and DL models, performs better [36]. BSFR-SH is a strong blockchain-supported security framework efficient for ransomware detection, mitigation, and data recovery in intelligent healthcare systems [37]. Techniques developed on ML have been a success for ransomware pattern identification, and RF, SVM, and NNs are capable of identifying with up to 97.3% accuracy. Honeypotting methods also offer passive monitoring functionality through decoy attraction of ransomware in isolated areas for monitoring purposes. Statistical anomaly detection methods have also been applied [38]. Network traffic monitoring can also identify the attacker's IP address. Comparison of various methods utilized for detecting ransomware, such as signature-based, ML-based, and dynamic analysis, are also presented in the work [39] ML-based approaches proved to be an effective solution that provides real-time detection of ransomware. Several prevention techniques that involve firewalls, intrusion detection systems, and regular backup products. The proposed system in this study is an extension of these existing methods by analyzing encrypted file extensions to identify ransomware type and recommend decryption tools [40].Using Gradient Tree Boosting, the detection improves to 99.997% accuracy with an FPR of just 0.01%. Static analysis, when coupled with ML techniques, is able to detect ransomware before execution effectively with reduced system damage potential [41]. The research also compares common ML algorithms with the comment that while age-old techniques such as SVM and RF are still in the forefront, newer DL techniques such as LSTM networks and CNNs have shown high potential to extract features independently and enhance detection performance [42]. Table III summarizes the ransomware detection ML and DL models.

TABLE III. RANSOMWARE DETECTION MODELS (ML AND DL)

Ref	Vulnerabilities / Challenges	Technology Used in Prevention	Year
[1]	Ransomware lifecycle exploitation via Windows APIs	API-based detection with ML models	2021
[25]	Accurate classification of ransomware families	Dual-stage ML architecture using Ensemble + LightGBM	2024
[26]	Behavioral threats in network/system activities	SOM, RF, LSTM models	2023
[27]	Zero-Day Ransomware Threats, Evasion Techniques, System Behavior Manipulation, Model Generalization	ML Algorithms: RF, SVM, and NNs.	2024
[28]	Fast detection/classification of ransomware	XGBoost and RF classifiers	2023
[29]	Advanced ransomware evading antivirus detection	Experimental ransomware and prevention insights	2021
[30]	Slow detection due to disassembly and analysis	Static analysis with RF, Gain Ratio	2020
[31]	Understanding ransomware paths	DNA act-Ran using ML and Digital DNA Sequencing	2020
[32]	Healthcare ransomware attacks	NBs classifiers with 99.2% accuracy	2024
[33]	Detection techniques categorization	Cyber Threat Hunting and dataset review	2022
[34]	Limitations of traditional detection	AI-based models (ML, DL, Ensemble Learning)	2024

[35]	Inefficient backup, detection, and communication	Blockchain cybersecurity models (Smart Contracts)	2024
[36]	Inconsistent malware detection	Hybrid of signature, behavior, DL models	2020
[37]	Smart healthcare ransomware attacks	BSFR-SH: Blockchain framework	2022
[38]	Diverse ransomware types (crypto, locker, etc.)	ML models (RF, SVM, and NN), Honeypots, Anomaly detection	2023
[39]	Crypto ransomware via phishing and fake updates	Forensic detection, traffic analysis, ML	2020
[40]	Encrypted data attacks demanding ransom	AI/ML for detection and auto-decryption tool suggestions	2023
[41]	High false positive rate in static analysis	Gradient Tree Boosting with 99.997% accuracy	2020
[42]	Advanced RaaS, double extortion, evasion	ML taxonomy (SVM, RF, LSTM, CNN) with benchmark analysis	2024

V. Analysis and Discussions

Using specific inclusion and exclusion criteria, the research methodology identified and evaluated malware detection models in accordance with a systematic literature review (SLR) approach. This guaranteed the results' objectivity, transparency, and reproducibility.

The file under review emphasizes the increasing relevance of ML and DL methods in ransomware detection in cloud computing and industrial IoT settings Conventional detection measures are not effective against contemporary ransomware because of polymorphism obfuscation and zero day attacks AI based measures are increasingly employed for sound and dynamic detection. ML models such as RF, SVMs, NBs and DTs remain in favor due to simplicity interpretability and efficiency RF model is most conventional to handle high dimensional data Ensemble methods such as lightGBM and boosting are more accurate, but ml relies heavily on handcrafted features and can be beaten by evasive ransomware. Deep models such as CNNs, RNNs, LSTM, Bidirectional LSTM (BiLSTM) models are gaining popularity These automatically extract complex patterns and temporal features from unprocessed data Autoencoders find use in anomaly detection DL achieves high accuracy at high frequency often over 99 but is plagued with problems of interpretability scalability and computational cost. Compared to ML model is simple and effective while DL works outstandingly well in classifying advanced evasive ransomware Ensemble and hybrid approaches yield best results incorporating efficiency and accuracy The use of blockchain federated learning and anomaly detection mechanisms is validated through research for enhancing resilience. Fig. 3 shows the classification methods of ML Approaches.

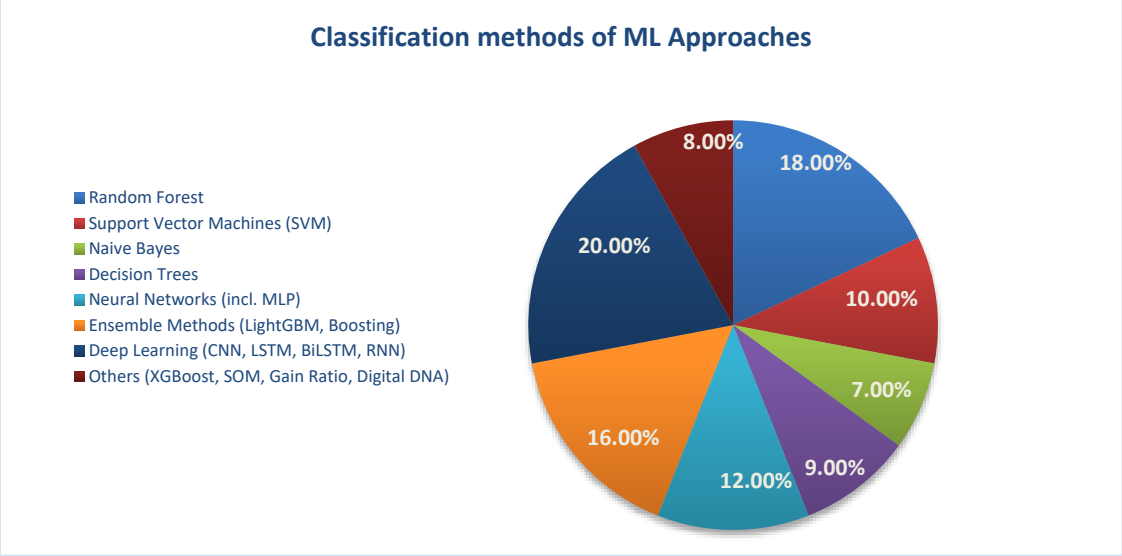


Fig. 3. Classification methods of ML approaches.

The most used methods are RFs and Ensemble Methods (18.00% each) due to their accuracy and robustness.

They are followed by NNs (12.00%), while classic models like SVM, NBs, and DTs still occupy an important role. hybrid approaches and neural solutions are gaining ubiquity, while traditional methods remain in common use for environments. Fig. 4 shows the classification methods of DL Approaches.

The most commonly used DL methods are RNNs, LSTM, BiLSTM, GRU at 40%, followed by CNNs at 30%. Hybrid CNN-RNN models take up 20%, which reflects the importance of architectural blends, and Autoencoders are utilized the least with 10%. In general, RNNs dominate due to their strength in sequence data handling, with CNNs and hybrids providing strong complementary techniques to ransomware detection. Fig. 5 shows the classification methods of ML approaches during past five years.

RF model was the leader in 2022 with nearly 100%, followed by the most commonly used method. After 2023, diversity picked up pace with XGBoost, Gradient Boosting, and SOM hitting about 30–35%.

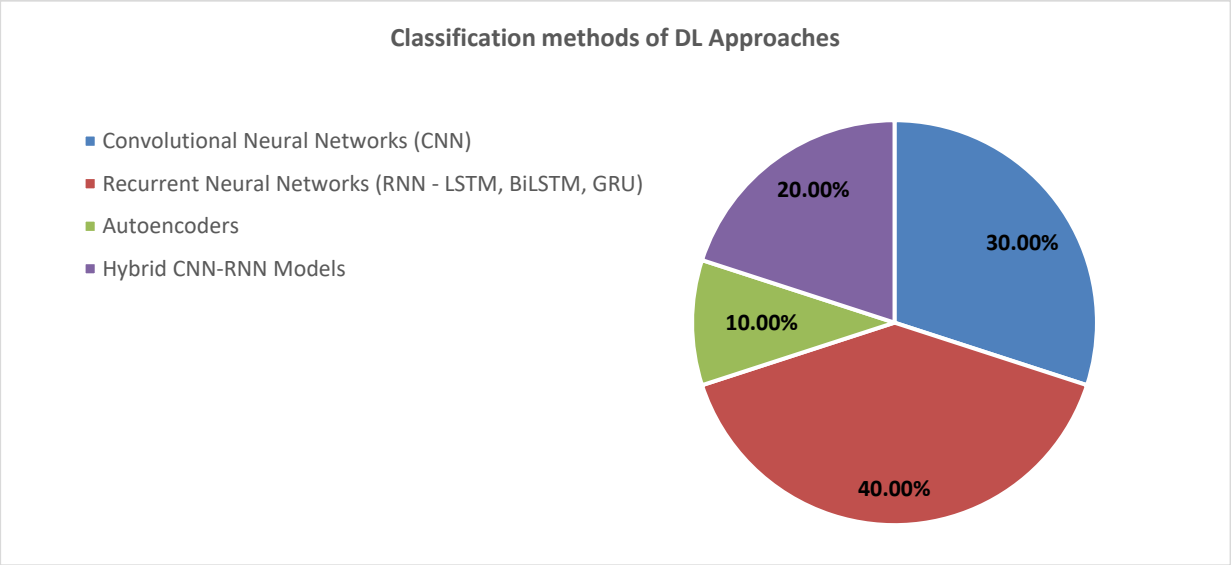


Fig. 4. Classification methods of DL approaches.

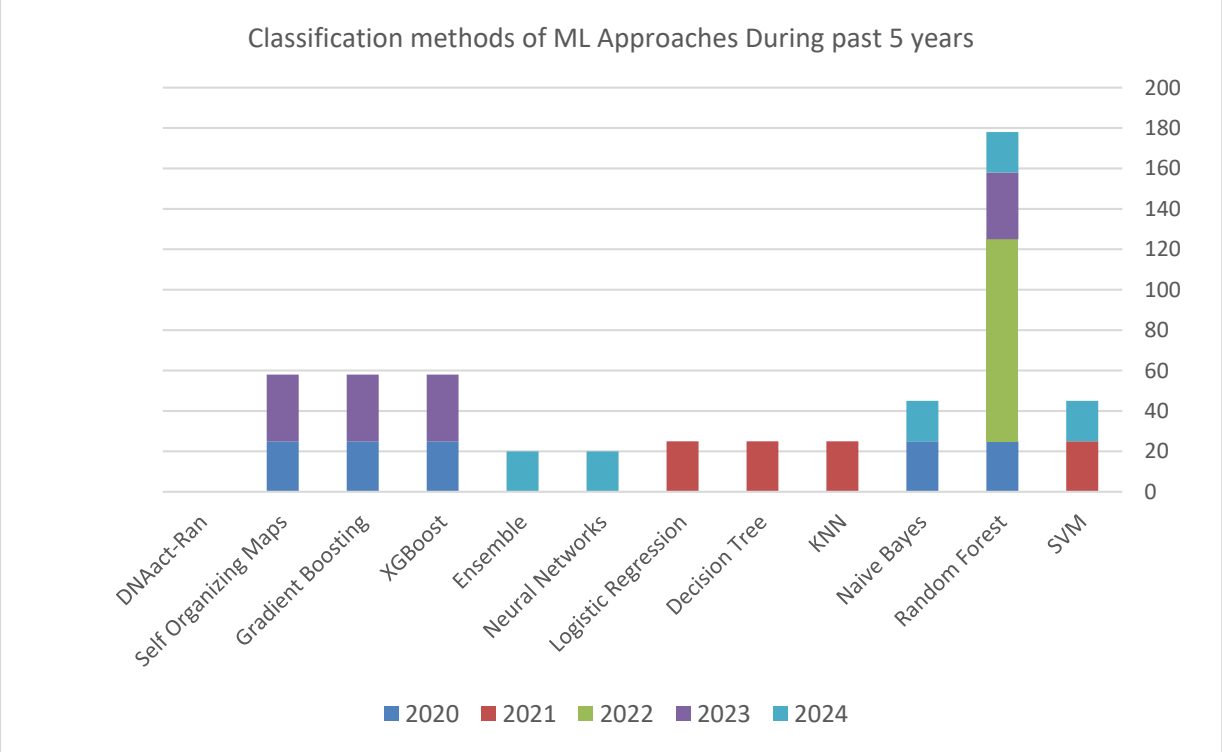


Fig. 5. Classification methods of ML approaches during past five years.

By 2024, NNs and Ensemble methods were popular (20%), which marked a shift towards complex and hybrid models.

- NSL-KDD has the maximum at 55%, the most sought-after dataset in all years.
- UNSW-NB15 leads with 28%, demonstrating its greater usage as a modern norm.
- Less frequently occurring data sets such as KDDCUP99 (6%), CIC-DoS2019 (5%), and CICIDS2017 (4%) are nonetheless useful for diversity.
- CICIDS2018 and ISCX2012 are used the least (1% each), reflecting minimal usage.

Overall, the findings reflect a continued overdependence on NSL-KDD and UNSW-NB15 with modest but useful contributions from newer datasets. Fig. 6 shows the different datasets for classification methods.

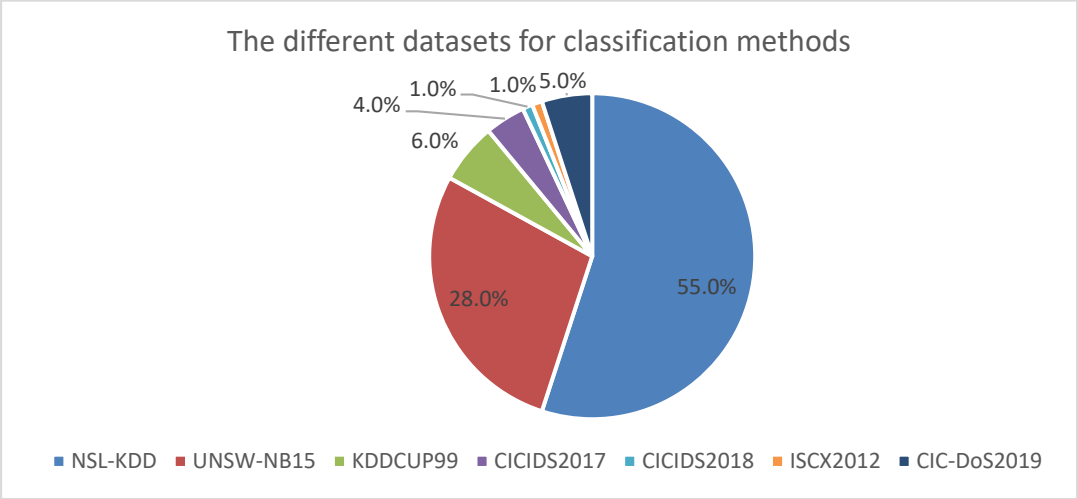


Fig. 6. The different datasets for classification methods.

VI. Conclusion

This work offers cybersecurity practitioners practical implications in addition to a comparative understanding of ransomware detection techniques. The results can help with the design of adaptive, hybrid security architectures that integrate blockchain, ML, and DL for practical protection systems. The findings from this research support that it is not one procedure that would suffice to offset the growing intricacy of cyber-attacks. Standard ML methods such as RF and Ensemble Techniques demonstrate good robustness and accuracy, while NNs provide greater adaptability. In contrast, DL approaches, specifically RNNs and CNNs, show improved accuracy in handling sophisticated and high-level threats, with hybrid CNN-RNN architectures further boosting detection. These results showcase the potential for constructing the future of resilient defense systems through an integration of hybrid and ensemble techniques, supported by blockchain-secured federated learning and adaptive anomaly detection. That type of integrated approach ensures a robust, scalable, and intelligent model for meeting current and future cyber threats.

Acknowledgment

The authors are thankful to the Deanship of Graduate Studies and Scientific Research at University of Bisha for supporting this work through the Fast-Track Research Support Program.

References

[1] Almousa, M., Basavaraju, S., and Anwar, M. (2021). Api-based ransomware detection using machine learning-based threat detection models. In *Proc. 18th Int.Conf. on Privacy, Security and Trust (PST), Dec. 2021*, pp. 1-7.

[2] Zhao, X. P., and Jiang, R. (2020). Distributed machine learning oriented data integrity verification scheme in cloud computing environment. *IEEE Access*, *8*, 26372-26384.

[3] Jamil, M. N., Schelén, O., Monrat, A. A., and Andersson, K. (2024). Enabling Industrial Internet of Things by Leveraging Distributed Edge-to-Cloud Computing: Challenges and Opportunities. *IEEE Access*.

[4] Sasikumar, K., and Nagarajan, S. (2024). Comprehensive review and analysis of cryptography techniques in cloud computing. *IEEE Access*.

[5] Hataba, M., Sherif, A., & Elkhouly, R. (2022). Enhanced obfuscation for software protection in autonomous vehicular cloud computing platforms. *IEEE Access*, *10*, 33943-33953.

[6] Gupta, I., Singh, A. K., Lee, C. N., and Buyya, R. (2022). Secure data storage and sharing techniques for data protection in cloud environments: A systematic review, analysis, and future directions. *IEEE Access*, *10*, 71247-71277.

[7] Miao, Y., Huang, Q., Xiao, M., and Li, H. (2020). Decentralized and privacy-preserving public auditing for cloud storage based on blockchain. *IEEE Access*, *8*, 139813-139826.

[8] Du, J., Dong, G., Ning, J., Xu, Z., and Yang, R. (2023). A Blockchain-Assisted Certificateless Public Cloud Data Integrity Auditing Scheme. *IEEE Access*, *11*, 123018-123029.

[9] Mann, Z. Á., Kunz, F., Laufer, J., Bellendorf, J., Metzger, A., and Pohl, K. (2021). RADAR: Data protection in cloud-based computer systems at run time. *IEEE Access*, *9*, 70816-70842.

[10] Aslan, Ö., Ozkan-Okay, M., and Gupta, D. (2021). Intelligent behavior-based malware detection system on cloud computing environment. *IEEE Access*, *9*, 83252-83271.

[11] Rao, S. M., and Jain, A. (2024). Advances in Malware Analysis and Detection in Cloud Computing Environments: A Review. *International Journal of Safety and Security Engineering*, *14*(1).

[12] Kimmel, J. C., Mcdole, A. D., Abdelsalam, M., Gupta, M., and Sandhu, R. (2021). Recurrent neural networks based online behavioural malware detection techniques for cloud infrastructure. *IEEE Access*, *9*, 68066-68080.

[13] Mofidi, F., Hounsinou, S. G., and Bloom, G. (2024, January). L-IDS: A multi-layered approach to ransomware detection in IoT. In *2024 IEEE 14th Annual Computing and Communication Workshop and Conference (CCWC)* (pp. 0387-0396). IEEE.

[14] Jawed, M. S., and Sajid, M. (2024, January). A Swarm Intelligence-based Faster and Secure Algorithm for Improved Industrial IoT-Cloud Computing Communication. In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)* (pp. 1-5). IEEE.

[15] Mahrous, W. A., Farouk, M., and Darwish, S. M. (2021). An enhanced blockchain-based IoT digital forensics architecture using fuzzy hash. *IEEE Access*, *9*, 151327-151336.

[16] Qiu, T., Chi, J., Zhou, X., Ning, Z., Atiquzzaman, M., and Wu, D. O. (2020). Edge computing in industrial internet of things: Architecture, advances and challenges. *IEEE Communications Surveys & Tutorials*, *22*(4), 2462-2488.

[17] Al-Hawawreh, M., Sitnikova, E., and Aboutorab, N. (2021). Asynchronous peer-to-peer federated capability-based targeted ransomware detection model for industrial IoT. *IEEE Access*, *9*, 148738-148755.

[18] Karacayılmaz, G., and Artuner, H. (2024). A novel approach detection for IIoT attacks via artificial intelligence. *Cluster Computing*, *27*(8), 10467-10485.

[19] Inaam ul Haq, M., Li, Q., and Hou, J. (2022). Analyzing the research trends of IoT using topic modeling. *The Computer Journal*, *65*(10), 2589-2609.

[20] Diro, A., Reda, H., Chilamkurti, N., Mahmood, A., Zaman, N., and Nam, Y. (2020). Lightweight authenticated-encryption scheme for internet of things based on publish-subscribe communication. *IEEE Access*, *8*, 60539-60551.

[21] Ullah, I., and Mahmoud, Q. H. (2021). Design and development of a deep learning-based model for anomaly detection in IoT networks. *IEEE Access*, *9*, 103906-103926.

[22] Ullah, I., and Mahmoud, Q. H. (2022). Design and development of RNN anomaly detection model for IoT networks. *IEEE Access*, *10*, 62722-62750.

[23] Saadat, H., Aboumadi, A., Mohamed, A., Erbad, A., and Guizani, M. (2021, June). Hierarchical federated learning for collaborative IDS in IoT applications. In *2021 10th Mediterranean Conference on Embedded Computing (MECO)* (pp. 1-6). IEEE.

[24] Humayun, M., Jhanjhi, N. Z., Alsayat, A., and Ponnusamy, V. (2021). Internet of things and ransomware: Evolution, mitigation and prevention. *Egyptian Informatics Journal*, *22*(1), 105-117.

[25] Yan, P., Khoei, T. T., Hyder, R. S., and Hyder, R. S. (2024, October). A Dual-Stage Ensemble Approach to Detect and Classify Ransomware Attacks. In *2024 IEEE 15th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 781-786). IEEE.

[26] Khurana, S. (2023, December). Ransomware Threat Detection and Mitigation using Machine Learning Models. In *2023 IEEE International Conference on ICT in Business Industry & Government (ICTBIG)* (pp. 1-6). IEEE.

[27] Brinkley, Y., Thompson, D., and Simmons, N. (2024). Machine learning-based intrusion detection for zero-day ransomware in unseen data.

[28] Kunku, K., Zaman, A. N. K., and Roy, K. (2023, December). Ransomware detection and classification using machine learning. In *2023 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 862-866). IEEE.

[29] Beaman, C., Barkworth, A., Akande, T. D., Hakak, S., and Khan, M. K. (2021). Ransomware: Recent advances, analysis, challenges and future research directions. *Computers & security*, *111*, 102490.

[30] Khammas, B. M. (2020). Ransomware detection using random forest technique. *ICT Express*, *6*(4), 325-331.

[31] Dutta, R., and Karmakar, S. (2024, June). Ransomware Detection in Healthcare Organizations using Supervised Learning Models: Naive Bayes Classifier. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-5). IEEE.

[32] Khan, F., Ncube, C., Ramasamy, L. K., Kadry, S., and Nam, Y. (2020). A digital DNA sequencing engine for ransomware detection using machine learning. *IEEE Access*, *8*, 119710-119719.

[33] Aldauiji, F., Batarfi, O., and Bayousef, M. (2022). Utilizing cyber threat hunting techniques to find ransomware attacks: A survey of the state of the art. *IEEE Access*, *10*, 61695-61706.

[34] Ferdous, J., Islam, R., Mahboubi, A., and Islam, M. Z. (2024). AI-based ransomware detection: A comprehensive review. *IEEE Access*.

[35] Fajri, A. I., Irawan, M. I., and Mahananto, F. (2024, August). A Systematic Literature Review on Blockchain-based Cybersecurity Models for Ransomware Mitigation. In *2024 IEEE International Symposium on Consumer Technology (ISCT)* (pp. 799-804). IEEE.

[36] Aslan, Ö. A., and Samet, R. (2020). A comprehensive review on malware detection approaches. *IEEE access*, *8*, 6249-6271.

[37] Wazid, M., Das, A. K., and Shetty, S. (2022). BSFR-SH: Blockchain-enabled security framework against ransomware attacks for smart healthcare. *IEEE Transactions on Consumer Electronics*, *69*(1), 18-28.

[38] Pandey, P., Jain, P., Gupta, A., Gupta, V. K., Jindal, H., and Gupta, A. (2023, November). Categorization, Detection, and Prevention of Ransomware Attack: A Review. In *2023 Seventh International Conference on Image Information Processing (ICIIP)* (pp. 552-557). IEEE.

[39] Ilker, K. A. R. A., and Aydos, M. (2020, October). Cyber fraud: Detection and analysis of the crypto-ransomware. In *2020 11th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)* (pp. 0764-0769). IEEE.

[40] Jayanthi, M. M., and Vijayakumar, K. (2023, May). Detection and decryption of ransomware. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 1264-1267). IEEE.

[41] Usharani, S., and Sandhya, S. G. (2020, July). Detection of ransomware in static analysis by using Gradient Tree Boosting Algorithm. In *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)* (pp. 1-5). IEEE.

[42] Ispahany, J., Islam, M. R., Islam, M. Z., and Khan, M. A. (2024). Ransomware detection using machine learning: A review, research limitations and future directions. *IEEE Access*.

A Comprehensive Survey of Visual SLAM Technology: Methods, Challenges, and Perspectives

Aidos Ibrayev, Amanzhol Bektemessov
International Engineering Technology University, Almaty, Kazakhstan

***Abstract*—Visual Simultaneous Localization and Mapping (Visual SLAM) has become a cornerstone of autonomous navigation and spatial understanding in robotics, augmented reality, and computer vision. This review presents a comprehensive examination of algorithmic progress in Visual SLAM, focusing on the three principal paradigms: monocular, stereo, and RGB-D SLAM. Monocular SLAM, known for its minimal hardware requirements, has evolved from feature-based methods to deep learning-enhanced systems, addressing challenges like scale ambiguity and drift. Stereo SLAM leverages depth through triangulation, improving scale accuracy and robustness, particularly in dynamic and low-texture environments. RGB-D SLAM, utilizing depth-sensing technology, has enabled dense and semantically enriched mapping, finding significant application in indoor and real-time scenarios. Through a chronological and technical exploration of representative methods including RatSLAM, ORB-SLAM, DSO, ProSLAM, ElasticFusion, DynaSLAM, and recent hybrid and learning-based frameworks. This review identifies major milestones and architectural innovations across paradigms. A cross-paradigm analysis highlights the trade-offs in accuracy, computational efficiency, and adaptability, while also discussing emerging trends such as semantic integration, multimodal fusion, and neural implicit representations. Furthermore, the paper outlines future directions that include lifelong learning, real-time deployment on edge devices, dynamic environment adaptation, and the convergence of geometry and learning-based pipelines. Supported by a detailed taxonomy and historical evolution illustrated in visual summaries, this review serves as a foundational reference for researchers and developers aiming to understand and contribute to the advancement of Visual SLAM technologies in both academic and real-world contexts.**

***Keywords*—Visual SLAM; monocular SLAM; Stereo SLAM; RGB-D SLAM; 3D mapping; pose estimation; loop closure; semantic SLAM; deep learning; sensor fusion**

I. Introduction

Simultaneous Localization and Mapping (SLAM) is a fundamental problem in robotics and computer vision, enabling an agent to build a map of an unknown environment while estimating its own pose within that space. Among the various modalities of SLAM, Visual SLAM (V-SLAM) has gained increasing attention due to its cost-effectiveness, wide applicability, and ability to capture rich scene semantics using standard cameras [1]. Unlike traditional SLAM approaches that rely on expensive range sensors such as LiDAR, visual SLAM uses images as the primary sensory input, offering a lightweight and scalable solution for many real-world applications including autonomous vehicles, augmented reality (AR), unmanned aerial vehicles (UAVs), and mobile robotics [2].

Visual SLAM systems are typically classified into three main categories based on the type of visual input: monocular, stereo, and RGB-D SLAM. Monocular SLAM uses a single camera and offers a compact and inexpensive setup but suffers from scale ambiguity and sensitivity to initialization [3]. Stereo SLAM employs a pair of cameras to estimate depth via triangulation, thereby overcoming scale issues and providing more stable estimates in textured environments [4]. RGB-D SLAM leverages cameras that deliver both color and depth information simultaneously, enhancing accuracy in indoor and structured scenes while introducing hardware constraints and limitations in outdoor or long-range settings [5].

Over the past two decades, the field has witnessed a remarkable evolution in algorithmic design, ranging from sparse, feature-based methods to dense and semi-dense approaches, and more recently, learning-based frameworks that integrate deep neural networks with traditional geometric pipelines [6]. Despite significant progress, many challenges persist, such as dealing with dynamic environments, achieving real-time performance on resource-constrained devices, and ensuring robustness under varying lighting or texture conditions [7]. Furthermore, achieving generalizable SLAM that can adapt across domains and scales remains an open research problem [8].

This review paper provides a structured and comprehensive overview of Visual SLAM methods across the three primary categories. By examining algorithmic trajectories, technical innovations, and representative benchmarks, this survey aims to map the evolving landscape of SLAM research and identify promising directions for future exploration.

II. Background and Fundamentals

Visual SLAM combines techniques from geometry, signal processing, and optimization to simultaneously estimate a camera's pose and reconstruct its environment. The core objective is to recover the trajectory $T_t = SE(3)$ and a map M from a sequence of image observations $\{I_t\}$ over time. Mathematically, the SLAM problem can be formulated as a maximum a posteriori (MAP) estimation:

$$\arg\max_{X,M} p(X, M \mid Z) \tag{1}$$

where, X is the set of poses, M is the map, and Z denotes the sensor measurements [9]. Factor graph models are

commonly used to express the probabilistic dependencies between variables in SLAM, allowing the estimation to be represented as a graph optimization problem [10].

A key task in visual SLAM is camera pose estimation, typically achieved by minimizing the reprojection error:

$$\min_{T} \sum_{i=1}^{N} \left\| u_i - \pi\left(T \cdot X_i\right) \right\|^2 \tag{2}$$

where, u_i are image features, π is the projection function, and X_i are 3D landmarks [11]. Map representation varies from sparse point clouds to dense voxel grids, depending on the method [12].

Loop closure detection corrects accumulated drift by identifying previously visited locations and enforcing pose graph consistency [13]. Pose graph optimization is often solved using nonlinear least squares:

$$\min_{X} \sum_{(i,j)\in\varepsilon}^{N} \left\| f_{ij}\left(X_i, X_j\right) - z_{ij} \right\|^2 \tag{3}$$

where, f_{ij} models relative motion and z_{ij} is the measured constraint [14]. Feature extraction, tracking, and data association form the visual front-end, while back-end optimization refines the trajectory and map [15].

Recent works also incorporate deep learning for depth estimation and semantic mapping, improving robustness in unstructured environments [16]. Benchmarks such as KITTI and TUM RGB-D provide standardized datasets for evaluation and comparison [17].

III. Algorithmic Paradigms for Slam

This section synthesizes the principal algorithmic paradigms that have shaped SLAM development for mobile robots, providing a structured overview of classical and contemporary frameworks. We first examine filter based methods such as EKF-SLAM and UKF-SLAM that leverage recursive Bayesian estimation to jointly infer robot trajectory and map features. Next, particle filter approaches and their Rao-Blackwellized variants (FastSLAM) are discussed for their ability to represent multimodal posteriors in highly nonlinear settings. We then turn to graph based optimization techniques, which cast SLAM as a sparse nonlinear least squares problem over pose and landmark variables, and explore incremental solvers such as iSAM. Finally, hybrid and emerging paradigms that integrate multiple estimation strategies, learning based components, and submapping schemes are reviewed to highlight the evolving landscape of SLAM algorithms.

A. Filter-Based Approaches

Filter based SLAM methods formulate the simultaneous estimation of robot pose and map features as a recursive Bayesian filtering problem [16]. At each time step, the joint state vector.

$$s_t = \begin{bmatrix} x_t \\ m \end{bmatrix} \tag{4}$$

where, x_t denotes the robot pose and m concatenates all landmark coordinates, is represented by a Gaussian belief $N\left(\hat{s}_t, P_t\right)$. The filter alternates between a prediction step

$$\begin{aligned} \hat{s}_{t|t-1} &= f\left(\hat{s}_{t-1}, u_t\right), \\ P_{t|t-1} &= F_t P_{t-1} F_t^T + Q_t \end{aligned} \tag{5}$$

and a correction step upon receiving a measurement z_t:

$$\begin{aligned} K_t &= P_{t|t-1} H_t^T \left(H_t P_{t|t-1} H_t^T + R_t\right)^{-1} \\ \hat{s}_t &= \hat{s}_{t|t-1} + K_t\left(z_t - h\left(\hat{s}_{t|t-1}\right)\right) \\ P_t &= \left(I - K_t H_t\right) P_{t|t-1} \end{aligned} \tag{6}$$

Here, f and h are the process and measurement models, F_t and H_t their Jacobians, and Q_t, R_t covariance matrices. Within this framework, the Extended Kalman Filter SLAM (EKF-SLAM) linearizes f and h about the current estimate, resulting in first-order approximations of non-linear dynamics and sensor models [17]. EKF-SLAM maintains a full covariance matrix across robot and landmark states, ensuring consistent uncertainty propagation but incurring $O\left(n^2\right)$ computational complexity in the number of landmarks n. To mitigate linearization errors, the Unscented Kalman Filter SLAM (UKF-SLAM) employs the unscented transform: a set of deterministically chosen sigma points $\left\{\chi_{t-1}^i\right\}$ is propagated through the true non-linear functions, yielding more accurate posterior mean and covariance estimates without explicit Jacobian computation [18]. UKF-SLAM typically achieves higher accuracy in highly non-linear regimes at the cost of increased per-step computation proportional to the number of sigma points ($2L+1$ for a state of dimension L) [19]. Both EKF-SLAM and UKF-SLAM serve as foundational benchmarks against which more advanced, optimization-based, and hybrid SLAM paradigms are compared.

B. Particle Filter SLAM

Particle-filter SLAM methods approximate the full SLAM posterior $p\left(x_{0:t}, m \mid z_{1:t}, u_{1:t}\right)$ by a finite set of N weighted particles $\left\{x_{0:t}^i, m^i, w_t^i\right\}_{i=1}^N$. At each time step, particles are propagated according to the motion model $x_t^i \approx p\left(x_t \mid x_{t-1}^i, u_t\right)$ and importance weights are updated via:

$$w_t^i \propto w_{t-1}^i, p\left(z_t \mid x_t^i, m^i\right) \tag{7}$$

Followed by resampling to focus computational resources on high-likelihood hypotheses. Unlike Gaussian-based filters, particle filters can represent arbitrary, multi-modal distributions, making them well-suited to highly non-linear and

ambiguous observation models. However, naive particle-filter SLAM suffers from the "curse of dimensionality," as the joint state space grows with the number of landmarks, leading to excessive computational demand and rapid sample impoverishment when N is insufficient.

FastSLAM Variants. FastSLAM addresses these challenges by factorizing the posterior into a product of a trajectory posterior and independent landmark posteriors:

Trajectory estimation is performed via a particle filter, while each landmark's position is maintained by a separate, low-dimensional Kalman filter (or alternative Gaussian estimator) [20]. FastSLAM 1.0 introduced this Rao-Blackwellized approach, reducing computational complexity from $O(NM^2)$ to $O(NM)$. FastSLAM 2.0 refined the proposal distribution by incorporating the current measurement into the sampling step, thereby improving sample efficiency and reducing particle deprivation in highly informative environments [21]. Subsequent variants have further enhanced robustness through adaptive resampling thresholds, hierarchical landmark clustering, and incorporation of grid-based mapping techniques for environments with dense feature distributions [22]. These innovations have established FastSLAM as a versatile, real-time SLAM framework on mobile-robot platforms with moderate computational resources.

C. Graph-Based Optimization

Graph-based SLAM reformulates the joint estimation problem as a nonlinear least-squares optimization over a sparse graph, in which nodes represent robot poses (and optionally landmark states) and edges encode spatial constraints derived from odometry and observations. In the pose-graph formulation, only robot poses $\{x_i\}$ are included as variables, and measurements between poses (e.g., odometry or loop closures) define binary factors. The optimization seeks

$$x^* = \arg\min_x \sum_{(i,j)\in\varepsilon} || z_{ij} \Theta(x_i^{-1} \oplus x_j) ||^2_{\Omega_{ij}} \quad (8)$$

where, z_{ij} is the measured relative transform between poses i and j, Ω_{ij} the information matrix, and $\oplus$, $\ominus$ manifold-specific composition and inverse-composition operators [23]. Factor-graph SLAM extends this concept by including landmark nodes $\{m_k\}$ and multiary factors, yielding a unified representation amenable to general-purpose solvers. The factor-graph cost function becomes.

$$\{x^*, m^*\} = \arg\min_{x,m} \sum_{f\in F} || \phi_f(x_f, m_f) - z_f ||^2_{\Omega_{ij}} \quad (9)$$

where, each factor f relates a subset of pose and landmark variables through the measurement function ϕ_f and information matrix Ω_f [24]. Sparsity in the graph structure enables efficient exploitation of sparse linear algebra techniques, yielding favorable scaling to large-scale environments.

Incremental smoothing and mapping (iSAM) algorithms tackle the computational demands of online graph optimization by updating the solution and factorization incrementally as new measurements arrive. The original iSAM algorithm incrementally constructs and maintains a square-root information matrix via QR or Cholesky factorization, avoiding the need to re-solve the entire system at each step [25]. iSAM2 further introduces a Bayes tree data structure that encapsulates the conditional dependency structure of variables, allowing selective relinearization and re-factorization only in affected subtrees when new factors are added or existing factors change [26]. This strategy achieves real-time performance on resource-limited platforms while preserving consistency and accuracy, making iSAM variants a de facto choice for optimization-based SLAM in mobile-robot applications.

IV. Visual SLAM

Visual SLAM (Simultaneous Localization and Mapping) is a technique that leverages visual input, typically from one or more cameras, to perform real-time mapping and localization in unknown environments. As shown in Fig. 1, Visual SLAM is broadly categorized into three main types: Monocular SLAM, Stereo SLAM, and RGB-D SLAM, each characterized by its unique camera configuration and depth perception capabilities. This classification allows for flexibility in application domains, ranging from lightweight mobile robotics to dense indoor reconstruction [18].

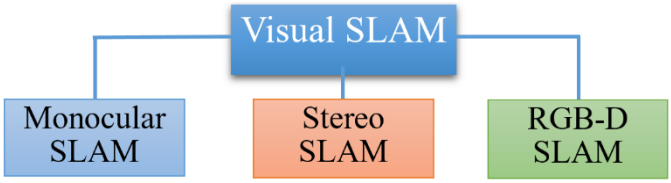


Fig. 1. Taxonomy of Visual SLAM: Classification into Monocular, Stereo, and RGB-D SLAM approaches.

Monocular SLAM employs a single camera and is attractive for its simplicity and low hardware cost. However, it faces challenges such as scale ambiguity and higher drift in long trajectories [19]. Stereo SLAM, by using two synchronized cameras, can triangulate depth directly, resulting in improved robustness and scale consistency [20]. In contrast, RGB-D SLAM integrates color and depth data from structured light or time-of-flight sensors, enabling dense and accurate reconstructions, particularly in indoor scenes [21].

Recent advancements in visual SLAM have focused on combining geometric methods with learning-based models to enhance robustness in dynamic or poorly textured environments [22]. Additionally, visual-inertial fusion and multi-sensor integration have further extended the capabilities of SLAM systems under challenging conditions [23]. These developments highlight the rapid evolution of visual SLAM technologies across all three categories.

V. Monocular SLAM: Evolution and Methods

Monocular SLAM has emerged as one of the most thoroughly investigated areas in the broader SLAM research landscape, primarily due to its minimal hardware requirements and its adaptability across diverse platforms and applications.

By using a single RGB camera, monocular SLAM systems aim to simultaneously estimate camera motion and reconstruct the surrounding environment. However, the absence of direct depth measurements introduces inherent challenges such as scale ambiguity, motion degeneracy in planar scenes, and increased sensitivity to lighting and texture variations [24]. As depicted in Fig. 2, the evolution of monocular SLAM has progressed through a series of methodological innovations in both the visual front-end and the back-end optimization strategies.

Initial efforts began with RatSLAM (2004), which was biologically inspired and modeled after rodent navigation systems. While simplistic, it laid foundational principles in appearance-based mapping. This was succeeded by MonoSLAM (2007), a pioneering approach that applied an Extended Kalman Filter (EKF) to jointly estimate camera pose and sparse 3D map points in real time [25]. The advent of PTAM in the same year introduced a game-changing architectural shift by decoupling tracking and mapping into parallel threads, drastically improving the performance of real-time monocular SLAM systems [26].

Subsequent methods diverged into two main categories: feature-based and direct methods. Feature-based approaches rely on the detection and matching of keypoints, while direct methods operate on image intensities without extracting features. DTAM and LSD-SLAM demonstrated that it is possible to achieve dense and semi-dense reconstructions using photometric information alone, resulting in more detailed environmental maps [27][28]. These techniques offered higher resolution outputs but required careful handling of photometric calibration and motion assumptions.

A major leap occurred with the release of ORB-SLAM in 2015, which combined ORB keypoint descriptors with a robust framework for tracking, mapping, loop closure, and relocalization [29]. Its success led to further extensions, including ORB-SLAM2 and ORB-SLAM3, which supported multiple sensor configurations such as stereo and RGB-D, and included inertial fusion for improved robustness in challenging motion conditions [30]. These improvements allowed for more accurate scale estimation and reduced drift during long-term operations [31].

Simultaneously, direct methods like DSO (Direct Sparse Odometry) and its variants, including Stereo DSO, LDSO, and V-LDSO, refined photometric error minimization and introduced keyframe-based filtering to enhance both accuracy and computational efficiency [32][33]. These methods, while sensitive to illumination changes, offered high precision in texture-rich environments and demonstrated robustness in frame-to-frame tracking without reliance on feature extraction.

Hybrid approaches have also gained attention, blending the strengths of direct and feature-based methods. Algorithms like SVO (Semi-Direct Visual Odometry), PL-SVO (Point and Line SVO), and CNN-SVO achieved a balance between robustness and efficiency by integrating direct photometric tracking with geometric landmarks [34][35]. These approaches reduced the dependency on handcrafted features and improved performance in a wider range of scene geometries.

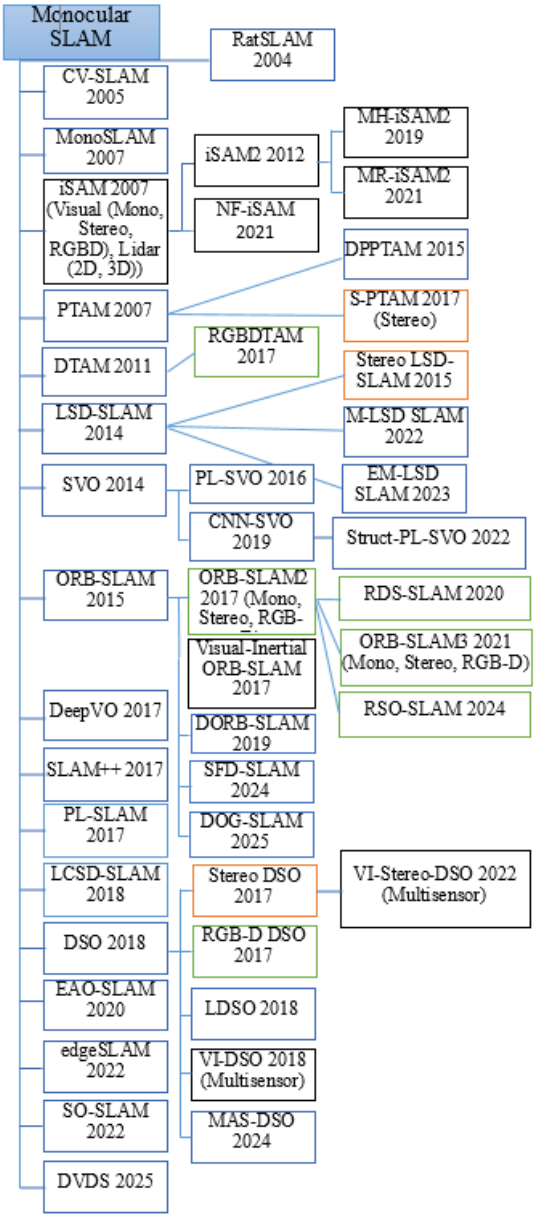


Fig. 2. Chronological development of monocular SLAM methods from 2004 to 2025.

In recent years, the incorporation of deep learning has significantly influenced the monocular SLAM domain. Frameworks such as DeepVO, edgeSLAM, and CNN-SLAM utilized neural networks for depth prediction, feature learning, and semantic scene understanding [36][37]. These models not only enhanced robustness in low-texture or dynamic environments but also enabled the inference of scene geometry from a single image. Moreover, methods such as SLAM++, SO-SLAM, and Struct-PL-SVO integrated object-level and structural cues into the mapping process, moving toward semantic and structural awareness in localization [38][39].

Recent state-of-the-art systems like DVDS (2025) and MAS-DSO demonstrate the growing emphasis on multi-sensor integration, semantic enrichment, and real-time adaptability [40][41]. These systems incorporate inertial data, semantic

segmentation, and learned depth priors to improve resilience in diverse and unpredictable environments.

Despite these advancements, monocular SLAM still faces unresolved issues, particularly in handling motion blur, rapid camera movements, dynamic object interference, and texture-poor scenes. The pursuit of robustness, generalization, and computational scalability continues to drive innovation in this domain, with future directions likely to focus on hybrid geometric-learning systems, unsupervised adaptation, and cloud-enhanced mapping frameworks.

VI. Stereo SLAM: Advances and Applications

Stereo SLAM systems utilize a pair of synchronized cameras to estimate depth via triangulation, enabling accurate 3D reconstruction and metric-scale motion estimation. This configuration eliminates the scale ambiguity inherent in monocular systems and enhances robustness in texture-poor or low-light environments [42]. As illustrated in Fig. 3, the development of stereo SLAM has progressed steadily with contributions that focus on improving accuracy, structural awareness, and adaptability in dynamic scenarios.

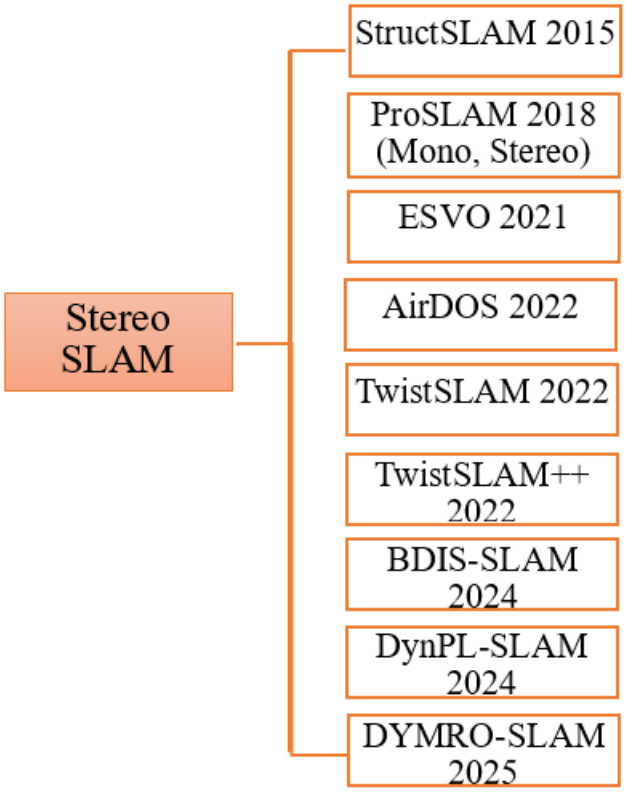


Fig. 3. Evolution of stereo SLAM methods and their applications.

StructSLAM (2015) laid the foundation by integrating stereo imagery with structural constraints to achieve semantically meaningful mapping [43]. The introduction of ProSLAM (2018) provided a lightweight and modular stereo SLAM framework capable of real-time performance, supporting both mono and stereo inputs for flexibility across platforms [44]. More recent methods, such as ESVO (2021) and AirDOS (2022), have extended stereo SLAM to address aerial applications and dynamic obstacle detection using stereo image alignment and depth filtering techniques [45][46].

The emergence of TwistSLAM and TwistSLAM++ (2022) represents a shift toward incorporating twist-based motion models for improved tracking under fast camera movements [47]. These systems enhance pose estimation stability while reducing drift in challenging trajectories. Building on these ideas, BDIS-SLAM and DynPL-SLAM (2024) have pushed stereo SLAM into the realm of dynamic environments, where the presence of moving objects demands adaptive strategies for scene segmentation and motion compensation [48].

Finally, DYMRO-SLAM (2025) exemplifies the future direction of stereo SLAM by integrating motion robustness, real-time segmentation, and adaptive feature association in highly dynamic and cluttered settings [49]. In addition to robotics, stereo SLAM has found applications in autonomous driving, drone navigation, and indoor mapping where depth accuracy and consistency are critical [50].

Collectively, the advancements in stereo SLAM reflect a balance between computational efficiency, structural integrity, and semantic awareness, making it a viable solution for diverse real-world deployment scenarios. However, the challenges of real-time performance in high-speed settings and long-term consistency in large-scale environments continue to motivate ongoing research.

VII. RGB-D SLAM: Depth-Enhanced Approaches

RGB-D SLAM systems integrate color (RGB) images with depth data, typically captured using structured light or time-of-flight sensors, to produce accurate and dense 3D maps in real-time. This additional depth channel enhances scene understanding, especially in indoor environments where scale accuracy and dense reconstruction are critical. As illustrated in Fig. 4, the evolution of RGB-D SLAM has been marked by diverse approaches that balance geometric precision, semantic richness, and computational efficiency.

Early RGB-D SLAM efforts began with KinectFusion (2011), which demonstrated the potential of consumer-grade depth sensors for dense real-time 3D reconstruction [51]. Shortly thereafter, Kintinuous (2012) extended KinectFusion by enabling continuous tracking and mapping across larger spaces [52]. Later methods such as ElasticFusion (2015) introduced surfel-based representations to maintain loop-closure consistency and surface flexibility during global optimization [53]. Similarly, BundleFusion (2017) incorporated global pose graph optimization with real-time feedback loops to minimize drift in long-term operations [54].

The year 2017 marked a proliferation of RGB-D SLAM algorithms. Systems such as VO-SF, Co-Fusion, and StaticFusion focused on improving tracking robustness and map representation in the presence of dynamic objects [55][56][57]. Others like MR-RGBD-SLAM and SWAIACP-SLAM added semantic cues and adaptive filtering mechanisms to improve depth consistency [58][59]. Simultaneously, methods like Detect-SLAM and CodeSLAM explored object-level SLAM and compact depth representations, respectively [60][61].

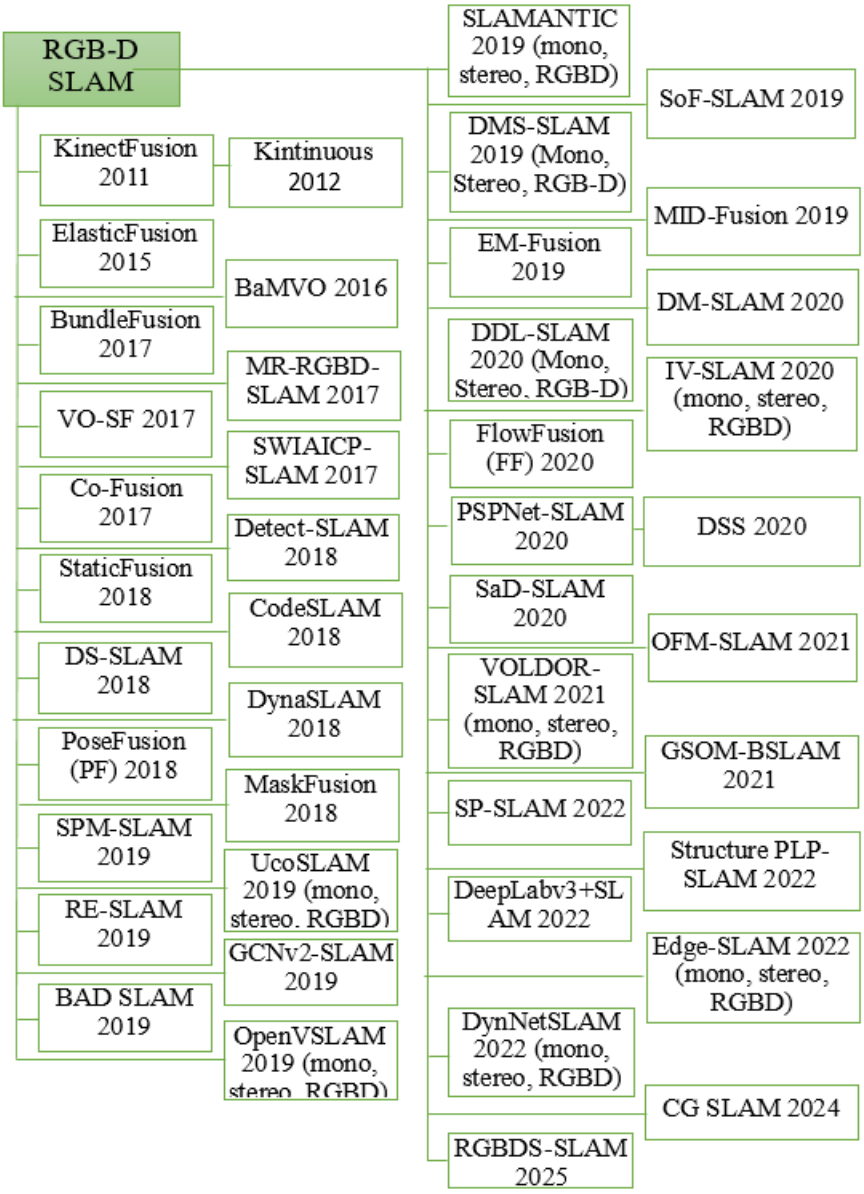


Fig. 4. Overview of RGB-D SLAM methods and their evolution.

In 2018, efforts like DS-SLAM, PoseFusion, and DynaSLAM focused on dynamic scene handling by introducing segmentation pipelines and learning-based object recognition to remove moving elements from mapping [62][63][64]. The inclusion of neural networks for motion segmentation and visual odometry marked a turning point, bridging traditional geometry-based techniques with deep learning paradigms [65]. Additionally, MaskFusion combined semantic instance segmentation with dense mapping, enabling object-level awareness in SLAM systems [66].

The year 2019 introduced generalized SLAM systems like UcoSLAM, OpenVSLAM, and SLAMANTIC, which supported multiple input modalities including mono, stereo, and RGB-D [67][68][69]. These frameworks emphasized modularity, scalability, and reusability, paving the way for flexible deployments in multi-sensor configurations. Simultaneously, RE-SLAM, SPM-SLAM, and BAD SLAM explored better loop closure techniques, keyframe management, and robustness under challenging illumination and texture conditions [70][71][72].

A wave of fusion-based systems emerged in 2019 and 2020, including EM-Fusion, DMS-SLAM, and DDL-SLAM, which aimed to fuse depth and color cues through probabilistic and optimization-driven pipelines [74][75]. FlowFusion, PSPNet-SLAM, and SaD-SLAM utilized optical flow and deep learning-based segmentation to enhance temporal consistency in fast-moving scenes [76][77][78]. Moreover, IV-SLAM and DM-SLAM extended these capabilities with multi-view consistency and dense depth propagation [79][80].

In 2021 and 2022, the emphasis shifted toward dynamic SLAM environments, as demonstrated by systems like OFM-SLAM, GSOM-BSLAM, and Structure PLP-SLAM, which integrated real-time semantic mapping and predictive modeling

of dynamic elements [81][82][83]. VOLDOR-SLAM and DynNetSLAM incorporated multi-branch neural networks for end-to-end depth estimation and pose prediction, effectively reducing dependency on handcrafted pipelines [84][85]. Additionally, Edge-SLAM brought attention to computational constraints, adapting RGB-D SLAM pipelines for edge devices by leveraging lightweight networks and compressed representations [86].

Recent approaches such as DeepLabv3+SLAM and SP-SLAM integrated semantic segmentation directly into the mapping process, enabling category-level understanding of environments [87][88]. This trend is furthered by hybrid methods like CG-SLAM and RGBDS-SLAM (2024–2025), which fuse cognitive perception models with conventional SLAM to support task-specific intelligence, such as scene classification and goal-directed navigation [89][90].

Notably, these RGB-D SLAM systems rely on benchmark datasets and real-time performance evaluation. Datasets like TUM RGB-D, ICL-NUIM, and ScanNet have been instrumental in validating performance across different benchmarks, offering standardized trajectories, dynamic scenes, and semantic annotations [91]. While these advancements have brought SLAM closer to robust deployment, issues related to sensor noise, occlusion, generalization in unseen environments, and energy efficiency still remain open challenges [92-94].

Overall, RGB-D SLAM has grown from simple depth-assisted mapping to rich, semantically-aware systems that integrate deep learning, object recognition, and multimodal fusion. This trajectory underscores the increasing convergence of geometric modeling and machine learning in the pursuit of highly adaptable and context-aware visual SLAM solutions.

VIII. Cross-Paradigm Analysis and Discussion

The evolution of Visual SLAM has produced a diverse array of systems across monocular, stereo, and RGB-D paradigms, each with specific strengths and limitations. A comparative analysis of these approaches reveals fundamental trade-offs in scalability, accuracy, computational complexity, and adaptability to dynamic environments.

Monocular SLAM remains the most lightweight and cost-effective option, requiring only a single camera for localization and mapping. However, it suffers from inherent scale ambiguity and limited depth perception, which restricts its application in scenarios demanding metric-scale accuracy [24]. Despite advancements like ORB-SLAM and DSO, monocular methods still rely heavily on loop closure and motion modeling to mitigate accumulated drift [29][32].

In contrast, Stereo SLAM inherently resolves the scale issue by estimating depth via triangulation. It provides a reliable middle ground between simplicity and depth accuracy, making it suitable for mobile robotics and drone navigation [42]. Systems such as ProSLAM and TwistSLAM++ demonstrate improved performance in fast-motion and low-texture environments [44][47]. Nevertheless, stereo systems require careful camera calibration and synchronization, which can introduce additional hardware constraints.

RGB-D SLAM, leveraging dense depth data from sensors like structured-light or time-of-flight cameras, excels in indoor mapping and 3D reconstruction [51][54]. These systems are particularly effective in dynamic and low-texture scenes due to the richness of depth information. However, they are limited by the operating range and sensitivity to lighting conditions, especially in outdoor settings [62]. Moreover, the high data throughput from RGB-D sensors increases computational demands, necessitating efficient data fusion and real-time optimization strategies [73][78].

Hybrid approaches integrating visual-inertial odometry, semantic segmentation, and learning-based modules are emerging across all three paradigms. For instance, visual-inertial extensions to ORB-SLAM and deep semantic mapping in RGB-D systems such as DynNetSLAM demonstrate that cross-domain fusion enhances robustness and scene understanding [30][85]. These trends reflect a shift toward generalizable SLAM systems capable of adapting to diverse environmental conditions and sensory inputs.

Despite these advances, unresolved challenges persist. Dynamic object handling, long-term consistency, real-time performance on edge devices, and scene generalization remain key bottlenecks. The convergence of geometry-based and data-driven methods is likely to play a pivotal role in addressing these issues and shaping the next generation of visual SLAM technologies.

IX. Future Directions

The field of Visual SLAM continues to experience rapid progress, yet numerous challenges remain unsolved, paving the way for future innovations. As the demand for intelligent autonomous systems rises, SLAM must evolve beyond static, structured environments to handle complex, unstructured, and dynamic real-world conditions. Several key directions are anticipated to define the next generation of SLAM systems.

One of the most pressing challenges is the robust handling of dynamic environments, where moving objects introduce inconsistencies in localization and mapping. While state-of-the-art methods such as DynaSLAM and DynNetSLAM have introduced dynamic object filtering using motion segmentation and learning-based modules, real-time robustness in highly dynamic scenarios remains limited [64][85]. Future systems must incorporate adaptive scene understanding and prediction mechanisms to account for object trajectories, occlusions, and interactions, thereby enhancing robustness under motion and clutter.

Another critical direction involves lifelong and continual SLAM, which allows systems to incrementally learn from their environments over extended periods. Current methods tend to operate in episodic modes, with limited memory and generalization across time or domains. Incorporating memory-augmented learning, self-supervision, and lifelong adaptation strategies would enable SLAM systems to refine their internal models continually and recover from localization failures or environmental changes. Techniques applied in DeepVO and other learning-based systems offer a foundation for such progress [36].

Semantic SLAM is also emerging as a transformative direction, where the integration of object-level understanding and scene semantics augments traditional geometric mapping. Systems such as MaskFusion, DeepLabv3+SLAM, and CG-SLAM demonstrate that embedding semantics enables higher-level reasoning, object manipulation, and contextual awareness [66][87][90]. Future research is expected to integrate multi-task learning pipelines that jointly optimize for pose estimation, semantic segmentation, and instance recognition. Such integration would be vital in service robotics, AR/VR, and autonomous navigation tasks where understanding scene content is as critical as knowing spatial geometry.

The development of resource-aware and edge-compatible SLAM algorithms is gaining significance as SLAM expands into mobile platforms, drones, and embedded robotics. High-performance SLAM often requires intensive computation and memory, making deployment on constrained devices difficult. Solutions like Edge-SLAM and optimized variants of SLAMANTIC and ORB-SLAM address these issues by incorporating lightweight architectures, neural compression, and on-device learning [86][69][30]. Future directions may involve further advances in model pruning, quantization, and hardware-aware design for real-time, low-power SLAM inference.

Multimodal SLAM represents another growing frontier. The fusion of multiple sensor inputs—including visual, inertial, depth, thermal, and even auditory data—enables robust mapping under diverse environmental conditions. Systems like ORB-SLAM3 and VOLDOR-SLAM illustrate how integrating inertial measurements and depth cues improves accuracy and consistency, especially in low-texture or fast-motion scenarios [30][84]. Expanding such systems to include semantic and audio signals could further enhance context-awareness, particularly in human-robot interaction and search-and-rescue missions.

In addition, learning-based SLAM continues to evolve. While geometry-based methods have dominated due to their interpretability and reliability, deep learning models are increasingly demonstrating their potential in monocular depth prediction, motion estimation, and loop closure detection. Hybrid pipelines that combine data-driven features with geometric consistency are likely to define future SLAM architectures. DMS-SLAM, CodeSLAM, and PSPNet-SLAM exemplify this convergence [74][61][77].

Finally, neural implicit representations, such as neural radiance fields (NeRF), are expected to redefine how SLAM systems model and store environments. These methods replace discrete maps with continuous, differentiable scene representations that offer high-fidelity reconstructions and efficient storage. Their integration with SLAM could lead to novel systems that unify localization, mapping, and scene rendering into a single framework, unlocking new possibilities in mixed reality and digital twin technologies.

In summary, the future of Visual SLAM lies in its ability to become more adaptive, intelligent, semantic, and efficient, leveraging cross-disciplinary advances in machine learning, sensor fusion, and cognitive modeling. These advancements will shape the next generation of autonomous systems capable of understanding, interacting with, and navigating the real world in an increasingly human-like manner.

X. Conclusion

This review has presented a structured and comprehensive analysis of the evolution, methodologies, and challenges in Visual SLAM across three primary paradigms: monocular, stereo, and RGB-D systems. Monocular SLAM remains valuable for its simplicity and affordability, yet it continues to face challenges related to scale ambiguity and robustness under complex motion. Stereo SLAM offers a balance between hardware complexity and depth accuracy, making it suitable for a wide range of robotic and autonomous applications. RGB-D SLAM systems have significantly advanced scene understanding and dense mapping capabilities, particularly in indoor and structured environments, though they remain constrained by sensor limitations and computational demands. Across all paradigms, recent trends reveal a growing convergence of traditional geometric methods with learning-based models, enabling SLAM systems to operate more robustly in dynamic and perceptually challenging scenarios. The incorporation of semantic information, multi-modal sensor fusion, and real-time optimization strategies has further extended the operational scope of SLAM technologies. Despite these advances, persistent issues such as adaptability in dynamic environments, efficient computation on edge devices, and generalization across unseen domains continue to hinder full autonomy. The future of Visual SLAM lies in the development of lifelong, context-aware systems capable of learning from experience, reasoning semantically, and operating reliably under uncertainty. By integrating insights from computer vision, robotics, and machine learning, the next generation of SLAM algorithms will move beyond spatial localization to support intelligent interaction and decision-making in real-world environments. This review serves as a foundation for understanding current capabilities and identifying key research directions in the evolving landscape of Visual SLAM.

Acknowledgment

This work was supported by the Science Committee of the Ministry of Higher Education and Science of the Republic of Kazakhstan within the framework of grant AP19679910 “Development of an experimental model of an autonomous unmanned mobile robot with off-road capability”, International Engineering Technological University, Almaty.

References

[1] T. Schneider et al., "Maplab: An Open Framework for Research in Visual-Inertial Mapping and Localization," in IEEE Robotics and Automation Letters, vol. 3, no. 3, pp. 1418-1425, July 2018, doi: 10.1109/LRA.2018.2800113.

[2] X. Zuo, P. Geneva, W. Lee, Y. Liu and G. Huang, "LIC-Fusion: LiDAR-Inertial-Camera Odometry," 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 2019, pp. 5848-5854, doi: 10.1109/IROS40897.2019.8967746.

[3] Y. -S. Shin and A. Kim, "Sparse Depth Enhanced Direct Thermal-Infrared SLAM Beyond the Visible Spectrum," in IEEE Robotics and Automation Letters, vol. 4, no. 3, pp. 2918-2925, July 2019, doi: 10.1109/LRA.2019.2923381

[4] T. Shan, B. Englot, C. Ratti and D. Rus, "LVI-SAM: Tightly-coupled Lidar-Visual-Inertial Odometry via Smoothing and Mapping," 2021

IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 2021, pp. 5692-5698, doi: 10.1109/ICRA48506.2021.9561996.

[5] Le, V.-H., Do, H.-S., Phan, V.-N., & Te, T.-H. (2024). TQU-SLAM Benchmark Feature-based Dataset for Building Monocular VO. Engineering, Technology & Applied Science Research, 14(4), 15330–15337. https://doi.org/10.48084/etasr.7611

[6] G. Ge, Y. Zhang, W. Wang, Q. Jiang, L. Hu, and Y. Wang, ‘‘Text-mcl: autonomous mobile robot localization in similar environment using text-level semantic information,’’ Machines, vol. 10, no. 3, Feb. 2022, 10.3390/machines10030169G. O. Young, “Synthetic structure of industrial plastics,” in Plastics, 2nd ed., vol. 3, J. Peters, Ed. New York, NY, USA: McGraw-Hill, 1964, pp.15–64.

[7] 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.

[8] S. Martínez-Díaz, “3D Distance Measurement from a Camera to a Mobile Vehicle Using Monocular Vision,” Journal of Sensors, vol. 2021, 5526931, 2021.

[9] S. Hensel, M. B. Marinov, M. Obert, “3D LiDAR Based SLAM System Evaluation with Low-Cost Real-Time Kinematics GPS Solution” Remote Sensing, Computation 10, no. 9, 154, 2022.

[10] 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

[11] Q. Lu, Y. Pan, L. Hu, and J. He, “A Method for Reconstructing Background from RGB-D SLAM in Indoor Dynamic Environments,” Sensors, vol. 23, no. 7, pp. 3529, Mar. 2023.

[12] F. Menna, A. Torresani, R. Battisti, E. Nocerino, and F. Remondino, “A Modular and Low-Cost Portable Vslam System for Real-Time 3d Mapping: From Indoor and Outdoor Spaces to Underwater Environments,” International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLVIII-2/W1-2022, pp. 153-162, 2022.

[13] I. Kalisperakis, T. Mandilaras, A. El Saer, P. Stamatopoulou, C. Stentoumis, S. Bourou, and L. Grammatikopoulos, “A Modular Mobile Mapping Platform for Complex Indoor and Outdoor Environments,” International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLIII-B1-2020, pp. 243-250, 2020.

[14] Z. Xie, Z. Li, Y. Zhang, J. Zhang, F. Liu, and W. Chen, “A Multi-Sensory Guidance System for the Visually Impaired Using YOLO and ORB-SLAM,” Information, vol. 13, no. 7, pp. 343, Jul. 2022.

[15] Omarov, B., Suliman, A., Kushibar, K. Face recognition using artificial neural networks in parallel architecture. Journal of Theoretical and Applied Information Technology 91 (2), pp. 238-248. Open Access.

[16] Alexander Pavlov, Kateryna Lishchuk, Oleg Melnikov, Mykyta Kyselov, Cennuo Hu, "Mathematics and Software for Coordinated Planning Using Aggregated Linear Volume-time Models of Discrete Manufacturing Systems", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.4, pp.1-15, 2025. DOI:10.5815/ijitcs.2025.04.01

[17] C. Mai, H. Chen, L. Zeng, Z. Li, G. Liu, Z. Qiao, Y. Qu, L. Li, and L. Li, “A Smart Cane Based on 2D LiDAR and RGB-D Camera Sensor-Realizing Navigation and Obstacle Recognition,” Sensors, vol. 24, no. 3, pp. 870, Jan. 2024.

[18] A. Altayeva, B. Omarov, H.C. Jeong, Y.I. Cho. Multi-step face recognition for improving face detection and recognition rate. Far East Journal of Electronics and Communications 16(3), pp. 471-491.

[19] D. Merkle and A. Reiterer, “Automated Method for SLAM Evaluation in GNSS-Denied Areas” Remote Sensing, vol. 15, no. 21, pp. 5141, 2023.

[20] J. Sun, J. Zhao, X. Hu, H. Gao, and J. Yu, “Autonomous Navigation System of Indoor Mobile Robots Using 2D Lidar”, Mathematics, vol. 11, no. 6, pp. 1455, 2023.

[21] Y. Wang, S. Zhang, and J. Wang, “Ceiling-View Semi-Direct Monocular Visual Odometry with Planar Constraint”, Remote Sens., vol. 14, no. 21, pp. 5447, 2022.

[22] V. Di Pietra, N. Grasso, M. Piras, and P. Dabove, “Characterization of a Mobile Mapping System for Seamless Navigation”, International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLIII-B1-2020, pp. 227–234, 2020.

[23] 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.

[24] D. Bolkas, M. O’Banion, and C. J. Belleman, “Combination of TLS and SLAM Lidar for Levee Monitoring”, ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. V-3-2022, pp. 641–647, 2022.

[25] L. He, Z. Jin, and Z. Gao, “De-Skewing LiDAR Scan for Refinement of Local Mapping”, Sensors, vol. 20, no. 7, pp. 1846, 2020.

[26] L. Moradi, M. Saadatseresht, and P. Shokrzadeh, “Development of a Voxel-Based Local Plane Fitting for Multi-Scale Registration of Sequential MLS Point Clouds,” ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. X-4/W1-2022, pp. 523–530, 2023.

[27] L. Chen, Z. Chen, and Z. Ji, “Expectation–Maximization-Based Simultaneous Localization and Mapping for Millimeter-Wave Communication Systems,” Sensors, vol. 22, no. 18, 6941, 2022.

[28] T. Bodrumlu and F. Caliskan, “Indoor Position Estimation Using Ultrasonic Beacon Sensors and Extended Kalman Filter”, Engineering Proceedings, vol. 27, no. 1, pp. 16, 2022. Presented at the 9th International Electronic Conference on Sensors and Applications, 1–15 November 2022.

[29] K. Xiao, W. Yu, W. Liu, F. Qu, and Z. Ma, “High-Precision SLAM Based on the Tight Coupling of Dual-Lidar Inertial Odometry for Multi-Scene Applications”, Applied Sciences, vol. 12, no. 3, 939, 2022.

[30] A. Basiri, V. Mariani, and L. Glielmo, “Improving Visual SLAM by Combining SVO and ORB-SLAM2 with a Complementary Filter to Enhance Indoor Mini-Drone Localization under Varying Conditions”, Drones, vol. 7, no. 6, 404, 2022.

[31] Al Noman, M. A., Zhai, L., Almukhtar, F. H., Rahaman, M. F., Omarov, B., 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. DOI: 10.11591/ijece.v13i1.pp347-357.

[32] Y. Zhai, B. Lu, W. Li, J. Xu, and S. Ma, “JD-SLAM: Joint camera pose estimation and moving object segmentation for simultaneous localization and mapping in dynamic scenes”, International Journal of Advanced Robotic Systems, vol. 18, no. 1, pp. 1–12, 2021.

[33] Y. Sun, J. Hu, J. Yun, Y. Liu, D. Bai, X. Liu, G. Zhao, G. Jiang, J. Kong, and B. Chen, “Multi-Objective Location and Mapping Based on Deep Learning and Visual SLAM,” Sensors, vol. 22, no. 19, 7576, 2022.

[34] Z. Chen, A. Xu, X. Sui, Y. Hao, C. Zhang, and Z. Shi, “NLOS Identification- and Correction-Focused Fusion of UWB and LiDAR-SLAM Based on Factor Graph Optimization for High-Precision Positioning with Reduced Drift”, Remote Sensing, vol. 14, no. 17, 4258, 2022.

[35] H. Wu, W. Wu, X. Qi, C. Wu, L. An, and R. Zhong, “Planar Constraint-Assisted LiDAR SLAM Algorithm Based on Manhattan World Assumption”, Remote Sensing, vol. 15, no. 1, 15, 2023.

[36] 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, 5457, 2022.

[37] G. Chen and L. Hong, “Research on Environment Perception System of Quadruped Robots Based on LiDAR and Vision”, Drones, vol. 7, no. 5, 329, 2023.

[38] J. Geng, A. A. Mokhtarzadeh, H. Fei, H. Gao, N. Ding, R. Kong, and J. Zhou, “Robot positioning and navigation technology is based on Integration of the Global Navigation Satellite System and real-time kinematics”, Journal of Physics: Conference Series, vol. 2467, no. 1, 012027, 2023.

[39] C. Xu, Z. Liu, and Z. Li, "Robust Visual-Inertial Navigation System for Low Precision Sensors under Indoor and Outdoor Environments", Remote Sensing, vol. 13, no. 4, 772, 2021.

[40] Q. Zhang and C. Li, "Semantic SLAM for mobile robots in dynamic environments based on visual camera sensors", Measurement Science and Technology, vol. 34, no. 8, 085202, 2023.

[41] Y. Xia, J. Cheng, X. Cai, S. Zhang, J. Zhu, and L. Zhu, "SLAM Back-End Optimization Algorithm Based on Vision Fusion IPS", Sensors, vol. 22, no. 23, 9362, 2022.

[42] O. F. Ince and J.-S. Kim, "TIMA SLAM: Tracking Independently and Mapping Altogether for an Uncalibrated Multi-Camera System", Sensors, vol. 21, no. 2, 409, 2021.

[43] C. Theodorou, V. Velisavljevic, and V. Dyo, "Visual SLAM for Dynamic Environments Based on Object Detection and Optical Flow for Dynamic Object Removal", Sensors, vol. 22, no. 19, 7553, Oct. 2022.

[44] S. Hensel, M. B. Marinov, M. Obert, "3D LiDAR Based SLAM System Evaluation with Low-Cost Real-Time Kinematics GPS Solution" Remote Sensing, Computation 10, no. 9, 154, 2022.

[45] H. Guan, C. Qian, T. Wu, X. Hu, F. Duan, and X. Ye, "A Dynamic Scene Vision SLAM Method Incorporating Object Detection and Object Characterization," Sustainability, vol. 15, no. 4, pp. 3048, Feb. 2023.

[46] P. Herbert, J. Wu, Z. Ji, and Y.-K. Lai, "Benchmarking Visual SLAM Methods in Mirror Environments," Computational Visual Media, vol. 10, no. 2, pp. 215-241, Apr. 2024.

[47] A. Elamin, N. Abdelaziz, and A. El-Rabbany, "A GNSS/INS/LiDAR Integration Scheme for UAV-Based Navigation in GNSS-Challenging Environments," Sensors, vol. 22, no. 24, 9908, 2022.

[48] Q. Lu, Y. Pan, L. Hu, and J. He, "A Method for Reconstructing Background from RGB-D SLAM in Indoor Dynamic Environments," Sensors, vol. 23, no. 7, pp. 3529, Mar. 2023.

[49] D. Chen, Q. Yan, Z. Zeng, J. Kang, and J. Zhou, "A Model of Real-Time Pose Estimation Fusing Camera and LiDAR in Simultaneous Localization and Mapping by a Geometric Method," Sensors and Materials, vol. 35, no. 1, pp. 167-181, 2023.

[50] F. Menna, A. Torresani, R. Battisti, E. Nocerino, and F. Remondino, "A Modular and Low-Cost Portable Vslam System for Real-Time 3d Mapping: From Indoor and Outdoor Spaces to Underwater Environments," International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLVIII-2/W1-2022, pp. 153-162, 2022.

[51] I. Kalisperakis, T. Mandilaras, A. El Saer, P. Stamatopoulou, C. Stentoumis, S. Bourou, and L. Grammatikopoulos, "A Modular Mobile Mapping Platform for Complex Indoor and Outdoor Environments," International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLIII-B1-2020, pp. 243-250, 2020.

[52] T. Lu, Y. Liu, Y. Yang, H. Wang, and X. Zhang, "A Monocular Visual Localization Algorithm for Large-Scale Indoor Environments through Matching a Prior Semantic Map" Electronics, vol. 11, no. 20, pp. 3396, 2022.

[53] W. Chen, G. Shang, K. Hu, C. Zhou, X. Wang, G. Fang, and A. Ji, "A Monocular-Visual SLAM System with Semantic and Optical-Flow Fusion for Indoor Dynamic Environments," Micromachines, vol. 13, no. 11, pp. 2006, Nov. 2022.

[54] Z. Xie, Z. Li, Y. Zhang, J. Zhang, F. Liu, and W. Chen, "A Multi-Sensory Guidance System for the Visually Impaired Using YOLO and ORB-SLAM," Information, vol. 13, no. 7, pp. 343, Jul. 2022.

[55] C. Bonfanti, G. Patrucco, S. Perri, G. Sammartano, and A. Spanò, "A New Indoor LiDAR-Based MMS Challenging Complex Architectural Environments," International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLVI-M-1-2021, pp. 79-86, 2021.

[56] T. Alhmiedat, A. M. Marei, W. Messoudi, S. Albelwi, A. Bushnag, Z. Bassfar, F. Alnajjar, and A. O. Elfaki, "A SLAM-Based Localization and Navigation System for Social Robots: The Pepper Robot Case," Machines, vol. 11, no. 2, pp. 158, Jan. 2023.

[57] C. Mai, H. Chen, L. Zeng, Z. Li, G. Liu, Z. Qiao, Y. Qu, L. Li, and L. Li, "A Smart Cane Based on 2D LiDAR and RGB-D Camera Sensor-Realizing Navigation and Obstacle Recognition," Sensors, vol. 24, no. 3, pp. 870, Jan. 2024.

[58] G. Patrucco, G. Sammartano, C. Bonfanti, and A. Spanò, "Assessing Terrestrial MMS 3D Data for Outdoor Multi-Scale Modelling," International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLVIII-1/W1-2023, pp. 371-378, 2023.

[59] W. Hu, K. Zhang, L. Shao, Q. Lin, Y. Hua, and J. Qin, "Clustering Denoising of 2D LiDAR Scanning in Indoor Environment Based on Keyframe Extraction", Sensors, vol. 23, no. 1, pp. 18, 2023.

[60] L. Morelli, F. Ioli, R. Beber, F. Menna, F. Remondino, and A. Vitti, "Colmap-Slam: A Framework for Visual Odometry", International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. XLVIII-1/W1-2023, pp. 317–324, 2023.

[61] D. Bolkas, M. O'Banion, and C. J. Belleman, "Combination of TLS and SLAM Lidar for Levee Monitoring", ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. V-3-2022, pp. 641–647, 2022.

[62] L. He, Z. Jin, and Z. Gao, "De-Skewing LiDAR Scan for Refinement of Local Mapping", Sensors, vol. 20, no. 7, pp. 1846, 2020.

[63] K. Jegadeeswari, R. Rathipriya, "Green AI Practices in Multi-objective Hyperparameter Optimization for Sustainable Machine Learning", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.2, pp.1-9, 2025. DOI:10.5815/ijitcs.2025.02.01

[64] L. Chen, Z. Chen, and Z. Ji, "Expectation–Maximization-Based Simultaneous Localization and Mapping for Millimeter-Wave Communication Systems," Sensors, vol. 22, no. 18, 6941, 2022.

[65] D. Damodaran, S. Mozaffari, S. Alirezaee, and M. J. Ahamed, "Experimental Analysis of the Behavior of Mirror-like Objects in LiDAR-Based Robot Navigation", Applied Sciences, vol. 13, no. 5, pp. 2908, 2023.

[66] Ł. Sobczak, K. Filus, J. Domańska, and A. Domański, "Finding the best hardware configuration for 2D SLAM in indoor environments via simulation based on Google Cartographer", Scientific Reports, vol. 12, pp. 18815, 2022.

[67] N. Li, L. Guan, Y. Gao, S. Du, M. Wu, X. Guang, and X. Cong, "Indoor and Outdoor Low-Cost Seamless Integrated Navigation System Based on the Integration of INS/GNSS/LIDAR System", Remote Sensing, vol. 12, no. 19, pp. 3271, 2020.

[68] T. Bodrumlu and F. Caliskan, "Indoor Position Estimation Using Ultrasonic Beacon Sensors and Extended Kalman Filter", Engineering Proceedings, vol. 27, no. 1, pp. 16, 2022. Presented at the 9th International Electronic Conference on Sensors and Applications, 1–15 November 2022.

[69] Wang, K., Guo, J., Chen, K., & Lu, J. (2025). An in-depth examination of SLAM methods: Challenges, advancements, and applications in complex scenes for autonomous driving. IEEE Transactions on Intelligent Transportation Systems.

[70] Wang, K., Zhao, G., & Lu, J. (2024). A deep analysis of visual SLAM methods for highly automated and autonomous vehicles in complex urban environment. IEEE Transactions on Intelligent Transportation Systems, 25(9), 10524-10541.

[71] Yan, C., Qu, D., Xu, D., Zhao, B., Wang, Z., Wang, D., & Li, X. (2024). Gs-slam: Dense visual slam with 3d gaussian splatting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 19595-19604).

[72] Zhu, S., Wang, G., Blum, H., Liu, J., Song, L., Pollefeys, M., & Wang, H. (2024). Sni-slam: Semantic neural implicit slam. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 21167-21177).

[73] 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.

[74] Lin, Z., Zhang, Q., Tian, Z., Yu, P., & Lan, J. (2024). DPL-SLAM: enhancing dynamic point-line SLAM through dense semantic methods. IEEE Sensors Journal, 24(9), 14596-14607.

[75] Zhang, H., Peng, J., & Yang, Q. (2024). PR-SLAM: parallel real-time dynamic SLAM method based on semantic segmentation. IEEE Access, 12, 36498-36514.

[76] He, L., Li, S., Qiu, J., & Zhang, C. (2024). DIO-SLAM: a dynamic RGB-D slam method combining instance segmentation and optical flow. Sensors (Basel, Switzerland), 24(18), 5929.

[77] Pan, H., Liu, D., Ren, J., Huang, T., & Yang, H. (2024). LiDAR-imu tightly-coupled SLAM method based on IEKF and loop closure detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 6986-7001.

[78] Islam, Q. U., Ibrahim, H., Chin, P. K., Lim, K., Abdullah, M. Z., & Khozaei, F. (2024). ARD-SLAM: Accurate and robust dynamic SLAM using dynamic object identification and improved multi-view geometrical approaches. Displays, 82, 102654.

[79] Rastorgueva-Foi, E., Kaltiokallio, O., Ge, Y., Turunen, M., Talvitie, J., Tan, B., ... & Valkama, M. (2024). Millimeter-wave radio SLAM: End-to-end processing methods and experimental validation. IEEE Journal on Selected Areas in Communications, 42(9), 2550-2567.

[80] Li, Y., Shen, H., Fu, Y., & Wang, K. (2024). A method of dense point cloud SLAM based on improved YOLOV8 and fused with ORB-SLAM3 to cope with dynamic environments. Expert Systems with Applications, 255, 124918.

[81] Jiang, J., Zhang, T., & Li, K. (2025). LiDAR-based 3D SLAM for autonomous navigation in stacked cage farming houses: An evaluation. Computers and Electronics in Agriculture, 230, 109885.

[82] Wang, Y., Ren, B., Zhang, X., Wang, P., Wang, C., Song, R., ... & Meng, M. Q. H. (2025). ROLO‐SLAM: rotation‐optimized LiDAR‐only SLAM in uneven terrain with ground vehicle. Journal of Field Robotics, 42(3), 880-902.

[83] Liang, M., Leitinger, E., & Meyer, F. (2025). Direct multipath-based SLAM. IEEE Transactions on Signal Processing.

[84] Tapu Biswas, Farhan Sadik Ferdous, Zinniya Taffannum Pritee, Akinul Islam Jony, "ScrumSpiral: An Improved Hybrid Software Development Model", International Journal of Information Technology and Computer Science(IJITCS), Vol.16, No.2, pp.57-65, 2024. DOI:10.5815/ijitcs.2024.02.05.

[85] Jiang, H., Xu, Y., Li, K., Feng, J., & Zhang, L. (2024). Rodyn-slam: Robust dynamic dense rgb-d slam with neural radiance fields. IEEE Robotics and Automation Letters.

[86] Huang, H., Li, L., Cheng, H., & Yeung, S. K. (2024). Photo-slam: Real-time simultaneous localization and photorealistic mapping for monocular stereo and rgb-d cameras. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 21584-21593).

[87] Zheng, J., Zhu, Z., Bieri, V., Pollefeys, M., Peng, S., & Armeni, I. (2025). Wildgs-slam: Monocular gaussian splatting slam in dynamic environments. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 11461-11471).

[88] Liu, X., Wen, S., Zhao, J., Qiu, T. Z., & Zhang, H. (2024). Edge-assisted multi-robot visual-inertial SLAM with efficient communication. IEEE Transactions on Automation Science and Engineering, 22, 2186-2198.

[89] Yang, Y., Pan, M., Tang, D., Wang, T., Yue, Y., Liu, T., & Fu, M. (2024). Mcov-slam: A multicamera omnidirectional visual slam system. IEEE/ASME Transactions on Mechatronics, 29(5), 3556-3567.

[90] Xin, Z., Yue, Y., Zhang, L., & Wu, C. (2024, May). Hero-slam: Hybrid enhanced robust optimization of neural slam. In 2024 IEEE International Conference on Robotics and Automation (ICRA) (pp. 8610-8616). IEEE.

[91] Yang, L., Ye, J., Zhang, Y., Wang, L., & Qiu, C. (2024). A semantic SLAM-based method for navigation and landing of UAVs in indoor environments. Knowledge-Based Systems, 293, 111693.

[92] Wen, S., Tao, S., Liu, X., Babiarz, A., & Yu, F. R. (2024). CD-SLAM: A real-time stereo visual–inertial SLAM for complex dynamic environments with semantic and geometric information. IEEE Transactions on Instrumentation and Measurement, 73, 1-8.

[93] Muhammad Amirul Asyraaf Roslan, Haryani Haron, "Designing the Smart Shopping Cart Mobile Application (SmartCart) Using Mobile Application Development Life Cycle", International Journal of Information Technology and Computer Science(IJITCS), Vol.16, No.4, pp.66-81, 2024. DOI:10.5815/ijitcs.2024.04.05.

[94] Wang, W., Wang, C., Liu, J., Su, X., Luo, B., & Zhang, C. (2024). HVL-SLAM: Hybrid vision and LiDAR fusion for SLAM. IEEE Transactions on Geoscience and Remote Sensing.

Uneven But Accelerating: AI Adoption in Higher Education

Mahendra Adhi Nugroho[1], Umar Yeni Suyanto[2], Didik Hariyanto[3], Septiningdyah Arianisari[4]
Department of Accounting Education Department-Faculty of Economics and Business,
Universitas Negeri Yogyakarta, Yogyakarta, Indonesia[1]
Department of Administration Education-Faculty of Economics and Business,
Universitas Negeri Yogyakarta, Yogyakarta, Indonesia[2]
Department of Electrical Engineering Education-Faculty of Engineering, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia[3]
Department of Business and Finance-Faculty of Vocational Studies, Universitas Negeri Yogyakarta, Yogyakarta, Indonesia[4]

***Abstract*—Artificial Intelligence (AI) is increasingly recognized as a transformative force in higher education, yet adoption remains patchy and often confined to partial implementations. Using the PRISMA protocol, this study systematically reviews 74 Scopus-indexed articles published between 2015 and 2025. Publication activity rose sharply after 2020, led by contributions from China, the United States, and Saudi Arabia. Across the corpus, Perceived Usefulness and the Technology Acceptance Model (TAM) are the most frequently applied constructs, while ethical and policy dimensions remain underexamined. Thematic analysis delineates five clusters: adaptive learning and personalization; ethics and trust; digital literacy and readiness; AI in assessment and evaluation; and organizational transformation. Despite growing attention, regional gaps persist—especially in developing countries, where constrained infrastructure, funding, and digital literacy impede adoption. To address these challenges, the study proposes a multi-level conceptual framework integrating TAM, UTAUT, TPACK, and TOE to connect individual, institutional, and external factors for sustainable AI-driven education. Overall, the review underscores that AI adoption is not merely an efficiency tool but a strategic lever to advance the Sustainable Development Goals (SDGs), particularly by fostering inclusive, equitable, and innovative higher education systems.**

Keywords*—Artificial intelligence adoption; higher education; sustainable education; developing country

I. Introduction

Artificial Intelligence (AI) has advanced rapidly and is now deployed across multiple domains, including education [1]. It is widely regarded as a promising instrument for improving educational quality over the long term [2]. In educational settings, AI can generate strategic efficiencies by delivering personalized feedback, automating administrative tasks, forecasting student performance, enhancing learner engagement, and enabling adaptive assessment [3]. Accordingly, further development and careful integration of AI in education are needed to realize comprehensive, system-wide benefits.

AI-driven innovations are reshaping higher education at multiple levels [4]. Beyond streamlining administrative processes, AI enables adaptive learning, personalized curricula, and automation of academic services [5] [6] [7]. Its adoption has been linked to improvements in teachers' psychological well-being—through workload reduction, heightened motivation, and strengthened assessment and evaluation skills [8]—while also promoting more personalized learning and richer student experiences [9]. Professional development has been shown to support effective integration of GenAI for enhancing student engagement in teaching practice [10] [11], an essential precursor to interaction and improved performance; accordingly, designing AI-enabled activities, including those using ChatGPT, is warranted. Evidence further suggests that AI can raise learning effectiveness by up to 30% via learning-style analytics and content-recommendation systems [12]. In the United States and China, AI-based chatbots already support interactive learning [13] [14]; in Serbia, artificial intelligence and machine learning have been used to cultivate collaborative learning environments, develop student skills, and provide accessible research opportunities [15]. In Australia, ChatGPT accounted for 17.3% of the variance in improved academic outcomes [16]. Overall, AI facilitates faster feedback for students and allows instructors to devote more attention to strategic, higher-order pedagogical work [17].

Adopting AI in higher education opens pathways to cultivate higher-quality student outcomes [18]. Implementation, however, remains uneven across countries. Hughes et. al. [19] report that GenAI uptake is hindered by heterogeneous knowledge of AI tools and processes and by underdeveloped GenAI regulation. In Nigeria, adoption is constrained by limited funding, scarce technical expertise, and inadequate infrastructure [20]. In China, low digital literacy and resistance to AI impede deeper integration [21]. Evidence from the Middle East and North Africa cites low State revenues and unclear policies as additional obstacles [22]. In Sri Lanka, the absence of university-level policies and guidelines further inhibits GenAI adoption [23]. Taken together, these findings underscore the diverse, context-specific challenges surrounding AI adoption in education. A fit-for-purpose development model is therefore critical as a strategic pathway to enable effective adoption across higher education institutions.

AI adoption models grounded in established frameworks such as UTAUT and TOE have been implemented in several countries [24] [25] [26]. However, these templates are not fully aligned with the realities of developing contexts marked by digital divides, institutional diversity, and distinct policy needs. Prior findings underscore the importance of determinants—government support, leadership, risk, strategy, system quality,

user satisfaction, and readiness—for information-system/technology adoption and the application of AI in learning [27] [30] [28] 29] [31] [32] [33] [34] 35] [36]. However, most studies remain confined to generic drivers of AI uptake and use. Recent evidence further indicates that ChatGPT adoption in higher education is shaped by Performance Expectancy, Effort Expectancy, Hedonic Motivation, and AI ethical issues (bias, privacy, transparency), which collectively call for educator training, AI literacy, critical-ethical competence, and practical learning (e.g., Moral Machine) to prepare an automation-ready workforce [37] [38] [39]. This stream of work, however, is limited to ChatGPT, lacks a holistic approach, and provides minimal discussion of AI adoption within local cultural and contextual settings.

Meanwhile, ChatGPT demonstrates fluency at the university level but falls short on critical, context-sensitive analysis. Teachers tend to view AI favorably, yet they often lack sufficient content knowledge [40] [41]. This study also has limitations in its analysis of student–lecturer readiness and in articulating an AI competency framework rooted in the local context. Mohd Rahim et. al. [42] propose a chatbot-based AI adoption model incorporating perceived trust, behavioral intention, and use among Malaysian graduate students; however, this study was constrained by a sample composed solely of university students. Alyousef et.al. [43] identify ease of use, usefulness, engagement, trust, familiarity, and behavioral intention as key drivers of AI adoption in higher education, but its reliance on a student-only sample limits generalizability beyond that context.

Celik [44] extends the TPACK framework by incorporating artificial intelligence and ethics, underscoring that teachers should master not only pedagogy, content, and technology but also ethical literacy in AI use. Its limitations include the lack of integration with other technology-adoption models and the omission of institutional and policy dimensions. Research [45] investigates factors influencing the adoption and use of AI-based voice assistants, but its household-only design yields a limited number of respondents. Rao and R. Prasad [46] examine the role of 5G technology as a leading enabler of Industry 4.0 and its integration with AI, the Internet of Things (IoT), big data, and cloud computing; however, it does not employ an adoption-theory framework, which is important for shaping AI acceptance at the institutional or individual level.

Okunlaya et. al. [47] examine how AI-enabled library services can catalyze digital transformation in higher education. Its principal limitation is the absence of a technology-adoption framework, leaving the determinants of AI acceptance unspecified. Huang and Lu [48] analyzes the implementation of AI in Intelligent Tutoring Systems (ITS) for continuing education but focus on developed-country contexts, leaving developing settings underexplored. More broadly, prior research has not integrated adoption models with institutional dimensions; consequently, AI acceptance—especially in developing countries—remains insufficiently understood.

Prior work has explored multiple facets of technology and AI adoption—government support, service quality, risk, strategy, system quality, user satisfaction, social capital, and readiness to adopt learning information technologies [27] [30] [28] 29] [31] [32] 33] [34] [35]—alongside applications of AI in teaching and learning [36]. Taken together, these studies tend to emphasize individual-level elements such as chatbots for academic services, AI-based learning analytics, and assessment automation. Nevertheless, a holistic model that integrates technological, organizational, environmental, policy, and human-resource readiness factors is still absent [37] [38] [39]. Moreover, most prior research overlooks the distinctive characteristics of higher education institutions—including infrastructure constraints, regulatory conditions, and digital readiness—thereby limiting contextual applicability [49] [40] [41].

The literature still reveals a gap in AI adoption within higher education. This study addresses that gap by conducting a systematic literature review of AI adoption using a systematic approach within the higher education ecosystem [50]. The review considers technological, organizational, human, environmental, and ease-of-adoption factors [51] [52]. By analyzing prior research, the study maps research distribution, synthesizes theoretical models, performs cluster analysis, identifies research gaps, derives implications, and proposes a conceptual framework for AI adoption in higher education. The aim is to provide insights attuned to technological, organizational, environmental, cultural, and policy conditions. The results are intended to guide future researchers in enhancing higher-education quality through AI adoption and to encourage policymakers to support sustainable educational transformation.

This study undertakes a systematic literature review to map the field, identify research trajectories and opportunities for development, and synthesize the factors that shape artificial intelligence adoption in higher education, thereby consolidating the evidence base in this domain. It seeks to address the following research questions:

RQ1: What is the state of higher education's readiness to adopt AI, and how is it related to continuing education?

RQ2: What is the potential of integrating cross-theoretical perspectives to build a conceptual framework of AI adoption for sustainable education?

RQ3: How can developing countries contribute to realizing AI-driven sustainable education?

RQ4: How does the conceptual framework of AI adoption support the goal of sustainable education?

The remainder of this article is organized as follows. Section II details the research methods, including the PRISMA protocol and the bibliometric and thematic procedures employed. Section III reports the analysis results on publication trends, dominant theoretical lenses, and the principal thematic clusters. Section IV concludes with the study's conclusions Finally, Section V discuss the study's limitations, and recommendations for future research.

II. Method

A. Research Design

This study adopts a qualitative design employing a systematic literature review. A systematic literature review systematically identifies studies on a defined topic to answer

research questions transparently while incorporating evidence and assessing study quality [53] [54]. It synthesizes strengths, weaknesses, and patterns of divergence across prior work to produce a comprehensive account and reveal research gaps [55]. We applied the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) protocol with Watase UAKE to identify systematic reviews [56], using PRISMA to ensure transparent reporting so that findings and procedures are fully described [57]. The review analyzes prior research on AI adoption models in higher education to provide a basis for developing adoption models aligned with educational conditions in Indonesia.

B. Identification

The literature search was conducted in Scopus using the keywords "AI Adoption," "Artificial Intelligence," and "Higher Education." This initial sweep identified 126 articles.

C. Screening

The retrieved records were screened against inclusion and exclusion criteria to ensure alignment with the study's aims. The inclusion criteria were: 1) publication in Scopus-indexed international journals (Q1–Q4), 2) a time window of 2015–2025, 4) empirical, developmental, or conceptual studies addressing AI adoption models in higher education, and 5) articles written in English. Applying these criteria led to the removal of four articles that fell outside the publication years and eighteen that were not Scopus Q1–Q4, leaving 106 articles at the screening stage.

A second pass applied the exclusion criteria: 1) articles discussing AI outside an educational context, 2) duplicate records, and 3) non–peer-reviewed items. Five articles met these exclusion criteria, yielding 101 articles.

D. Eligibility

Titles, abstracts, and full texts were then assessed for eligibility using the same criteria to confirm topical relevance. Of the 101 articles entering this phase, 30 failed to meet the eligibility threshold, leaving 71 articles.

E. Appraisal of Quality

A quality appraisal ensured substantive relevance to AI adoption in higher education. The researcher examined methodologies associated with the adoption models TAM, UATUT, TOE, DOI, and AITAM, and verified that studies addressed organizational, technological, policy, cultural, and environmental factors. All 71 articles from the prior step were deemed eligible. At this point, three additional articles judged suitable were added, bringing the final corpus to 74 articles. The following Fig. 1 is a chart of the article selection process:

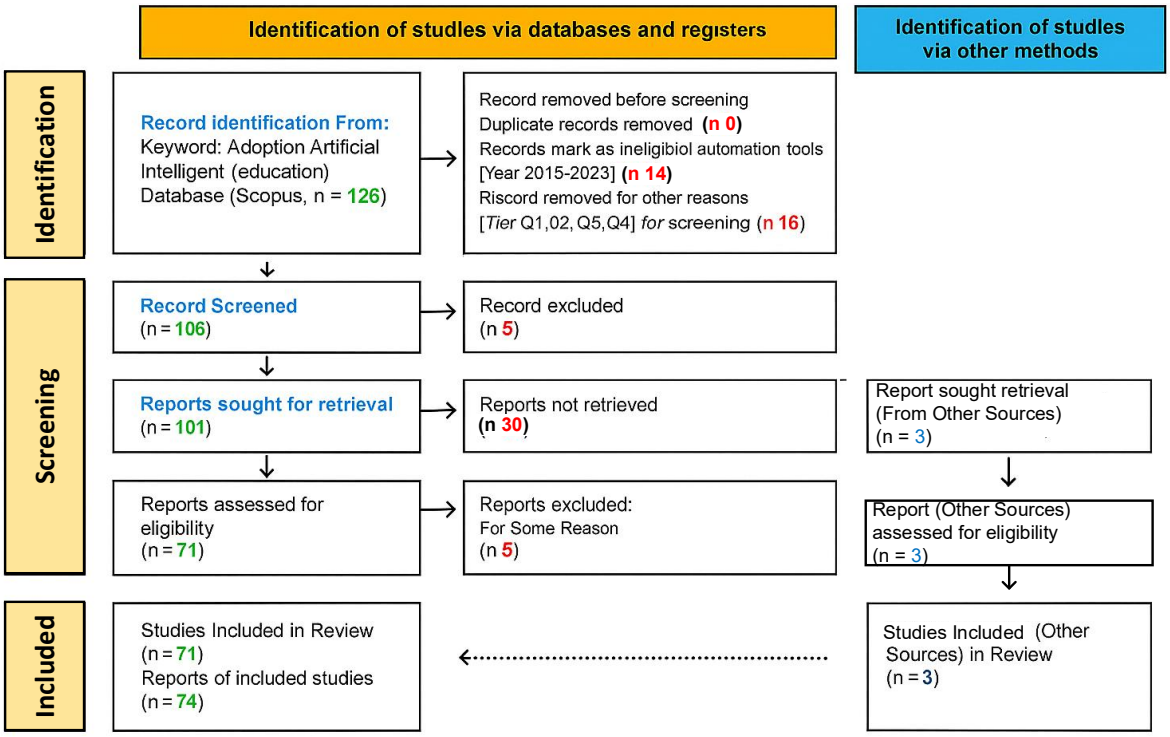


Fig. 1. Selection of articles in the PRISMA flow diagram.

III. Results and Discussion

This section reports findings from the systematic literature review. The discussion is organized into three parts: descriptive statistics, thematic analysis, and theoretical and practical implications.

A. Overview of Study Selection

Using the Scopus database and the keyword Adoption Artificial Intelligence Education for the 2015–2025 period, 126 articles were initially identified. Of these, 20 did not meet the criteria (outside the year range, not in the education category, or published in journals below the quality standard), leaving 106

for screening. The screening phase yielded 101 records that advanced to the retrieval stage. Subsequently, 71 articles satisfied the inclusion criteria, and three additional articles were added from other sources, producing a final corpus of 74 articles. This process ensured coverage of recent developments published in reputable international journals (Q1–Q4) and relevance to AI adoption in higher education.

B. Geographical Distribution of Studies

The country-level distribution reveals pronounced disparities. China led with 14 publications, followed by the United States (7) and Saudi Arabia (5). Indonesia and Malaysia contributed three articles each, while Mexico, Spain, Morocco, South Africa, and Turkey produced two each. China's dominance situates East Asia as a hub of AI-in-education research, influenced by substantial ed-tech investment and assertive government policy. The United States contributes notably to evaluative, pedagogical, and ethical strands, whereas Saudi Arabia's output aligns with national digitization priorities (Saudi Vision 2030). Contributions from Southeast Asia—including Indonesia and Malaysia—and from Africa are emerging but remain limited, underscoring a regional research gap that scholars in developing countries can address with locally grounded perspectives. A graph of article distribution by country is provided in Fig. 2.

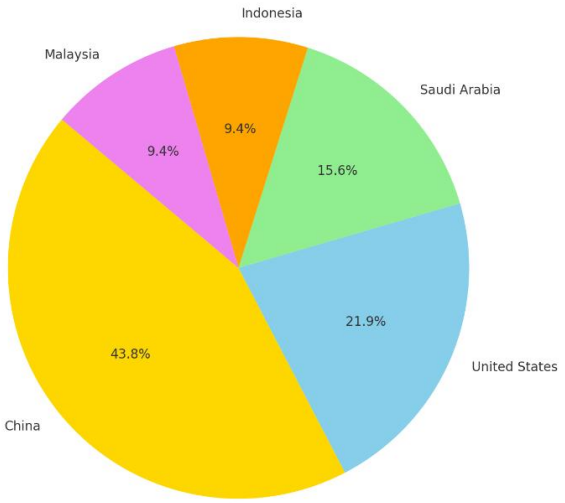


Fig. 2. Top 5 geographical distribution of studies.

C. Yearly Article Trends

Annual output shows a pronounced inflection after 2020. Between 2015 and 2020, publications were relatively flat at $\leq$ 2 articles per year. Starting in 2021, output rose sharply to 6 articles, followed by steady increases in 2022 and 2023 (11 articles each). The peak arrived in 2024 with 23 articles, signalling intensifying global attention to AI adoption in education. By mid-2025, 15 articles had already appeared, suggesting that the year's total may surpass prior records. This trajectory likely reflects exogenous drivers—most notably the COVID-19 pandemic's acceleration of educational digitization, policy support across countries, and expanded funding for AI-related research—pushing AI integration to the forefront of educational agendas in infrastructure, pedagogy, and ethics. A graph of article counts by year is presented in Fig. 3.

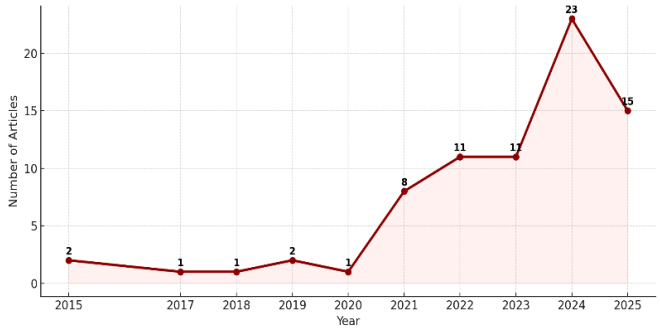


Fig. 3. Yearly article trends.

D. Variable Trends

The trend analysis indicates that Perceived Usefulness is the most frequently examined construct in AI - adoption research, appearing in 10 articles. This prominence suggests that the perceived benefits of technology are the dominant driver of readiness to adopt AI in higher education. Nine additional variables—Optimism, Perceived Privacy Risk, Social Influence, Artificial Intelligence, Effort Expectancy, Perceived Enjoyment, Trust, Perceived Ease of Use, and Perceived Ease—each appeared 2 times, underscoring the salience of social dimensions such as trust and usability in the adoption process. By contrast, Objective Useability, Insecurity, Relative Advantage, Compatibility, Expectation Confirmation, and Computer Playfulness appeared once each, signalling opportunities for deeper investigation of these factors in advancing sustainable education. The variable trends are presented in Fig. 4.

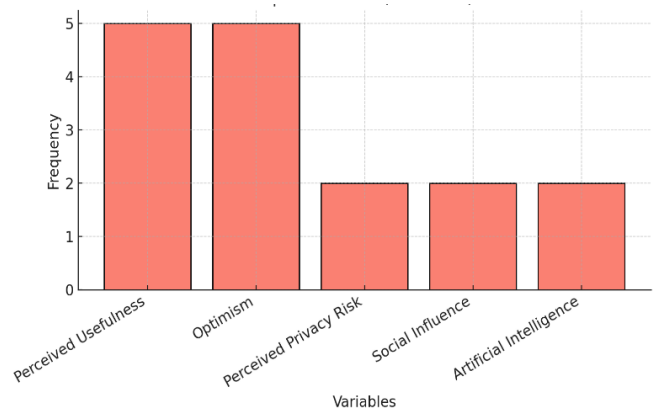


Fig. 4. Variable trends.

E. Theoretical Model of AI Adoption

Analysis of theoretical framing indicates that TAM (Technology Acceptance Model) is the most frequently employed approach, particularly for assessing faculty and student perceptions of AI's usefulness and ease of use. UTAUT/UTAUT2 appears in broader applications, incorporating performance expectancy, social influence, and facilitating conditions. At the institutional tier, TOE (Technology–Organization–Environment) is used to foreground organizational readiness, managerial support, and external pressures. At the same time, Diffusion of Innovation (DOI) is pertinent for explaining adoption dynamics across campus populations. Despite this range, most studies rely on a single

framework. The paucity of cross‐model integration is a notable limitation, given that AI adoption in higher education is intrinsically multidimensional—encompassing individual, institutional, and policy dimensions.

F. Thematic Clusters from Network Analysis

Keyword network analysis identified five principal clusters: 1) Adaptive Learning & Personalization—centered on intelligent tutoring systems, intelligent-TPACK, and AI-based learning personalization; 2) Ethics & Trust—highlighting transparency, student data privacy, and potential algorithmic bias; 3) Digital Literacy & Readiness—emphasizing faculty/student technology literacy and AI integration; 4) AI in Assessment & Evaluation—covering automated assessment, learning analytics, and academic performance prediction; and 5) Organizational Transformation—focusing on institutional readiness, change-management strategies, and digital-infrastructure investment. These findings indicate a fragmented landscape: clusters largely operate in parallel, with limited cross-theme integration.

The analysis further shows that AI technology constitutes the largest node, influencing performance, trust, organizational capabilities, and sustainability—evidence of topic dominance within AI acceptance. Major concentrations also appear across the pedagogy and literacy, ethics and sustainability, organization and performance, and technology adoption and acceptance clusters. Overall, research on AI adoption in higher education remains broad yet dispersed rather than fully integrated. Thematic clusters are presented in Fig. 5.

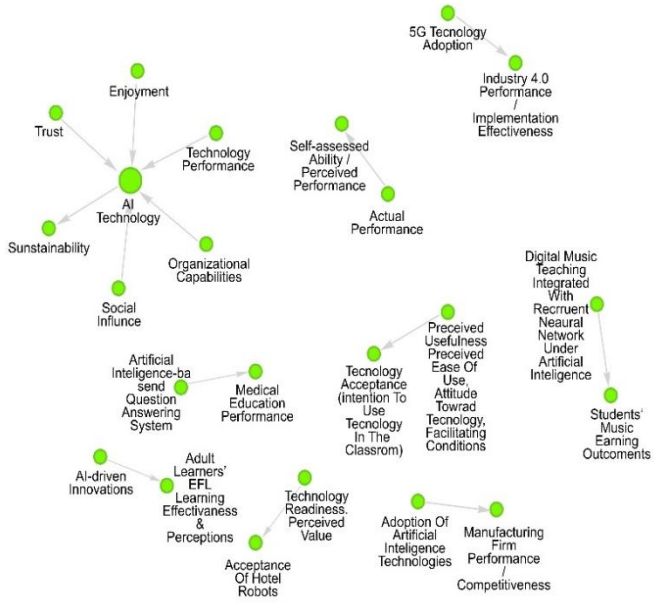


Fig. 5. Thematic clusters from network analysis.

G. Highly Cited Benchmark Articles

Citation analysis highlights several anchor studies. McLean & Osei-Frimpong (2019), with 536 citations, investigated determinants of adoption for AI-based virtual assistants (Alexa). Celik (2023), cited 330 times, advanced the concept of Intelligent (TPACK), embedding AI within teachers' pedagogical-competency frameworks. Lin et al. (2023), with 184 citations, emphasized AI's role in intelligent tutoring systems, corroborating its effectiveness for personalized learning. Wang et al. (2024), with 121 citations, offered a systematic review of AI in education that serves as a key roadmap for subsequent inquiries. Collectively, these benchmark articles illustrate that work combining strong theoretical innovation with clear practical relevance tends to attract broad scholarly attention. Table I depicts citations from the SLR analysis.

TABLE I. HIGHLY CITED BENCHMARK ARTICLES

Rank	Journal	Tier	Author	Year	Cites	Tittle
1	Computers in Human Behavior	Q1	McLean, Graeme; Osei-Frimpong, Kofi	2019	536	Hey Alexa, examine the variables influencing the use of artificial intelligence in in-home voice assistants.
2	Computers in Human Behavior	Q1	Celik, Ismail	2023	330	Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education
3	Wireless Personal Communications	Q3	Rao, Sriganesh K.; Prasad, Ramjee	2018	278	Impact of 5G Technologies on Industry 4.0
4	Smart Learning Environments	Q1	Lin, Chien-Chang; Huang, Anna Y. Q.; Lu, Owen H. T.	2023	184	Artificial intelligence in intelligent tutoring systems towards sustainable education: a systematic review
5	Library Hi Tech	Q2	Okunlaya, Rifqah Olufunmilayo; Syed Abdullah, Norris; Alias, Rose Alinda	2022	178	Artificial intelligence (AI) library services provide an innovative conceptual framework for the digital transformation of university education.

Rank	Journal	Tier	Author	Year	Cites	Tittle
6	Journal of Cases on Information Technology	Q4	Tahiru, Fati	2021	138	AI in Education
7	Expert Systems with Applications	Q1	Wang, Shan; Wang, Fang; Zhu, Zhen; Wang, Jingxuan; Tran, Tam; Du, Zhao	2024	121	Artificial intelligence in education: A systematic literature review
8	International Journal on Interactive Design and Manufacturing (IJIDeM)	Q3	Hernandez-de-Menendez, Marcela; Escobar Díaz, Carlos; Morales-Menendez, Ruben	2019	99	Technologies for the future of learning: state of the art
9	IEEE Access	Q2	Essa, Saadia Gutta; Celik, Turgay; Human-Hendricks, Nadia Emelia	2023	95	Personalized Adaptive Learning Technologies Based on Machine Learning Techniques to Identify Learning Styles: A Systematic Literature Review
10	Information	Q4	Sajja, Ramteja; Sermet, Yusuf; Cikmaz, Muhammed; Cwiertny, David; Demir, Ibrahim	2024	94	Artificial Intelligence-Enabled Intelligent Assistant for Personalized and Adaptive Learning in Higher Education

H. Research Gaps Identified

The synthesis reveals five persistent gaps. First, theoretical integration remains limited: studies continue to rely partially on TAM/UTAUT without adopting multi-level frameworks. Second, a regional skew persists, with East Asia dominating and developing countries contributing comparatively little. Third, longitudinal evidence is scarce; cross-sectional designs predominate, leaving long-term impacts largely unexamined. Fourth, ethical and policy dimensions—now increasingly salient in higher education—are underexplored. Fifth, links to academic outcomes are weak, as most research halts at adoption perceptions rather than connecting directly to student learning results.

I. Implications for Higher Education

Findings from this SLR carry multi-level implications. Academic: develop new conceptual frameworks that fuse technology-adoption theory with pedagogy (e.g., TAM + TPACK). Practical: universities should institute AI-literacy training for lecturers and students and cultivate an AI-enabled learning ecosystem. Policy: regulations on the ethical use of AI, student-data protection, and evaluation standards require prompt development. Global–local: best practices from China and the United States can be adapted, but must be calibrated to Indonesia's infrastructural realities and learning culture.

J. Conceptual Framework

Building on the findings, the framework positions individual factors (digital literacy, attitude, motivation), institutional factors (management support, infrastructure, campus policy), and external factors (regulation, global trends) as primary determinants of AI adoption in higher education. AI adoption, in turn, affects learning outcomes—namely personalization, effectiveness, and 21st-century literacy. Innovation can trigger insecurity, thereby motivating the development of intelligent TPACK; simultaneously, optimism and AI literacy shape perceived ease, AI usability, and trust. When AI is perceived as valuable, beneficial, and safe, acceptance rises. These dynamics foster positive attitudes that lead to intention and, ultimately, actual use. The AI adoption model derived from this review is presented in Fig. 6.

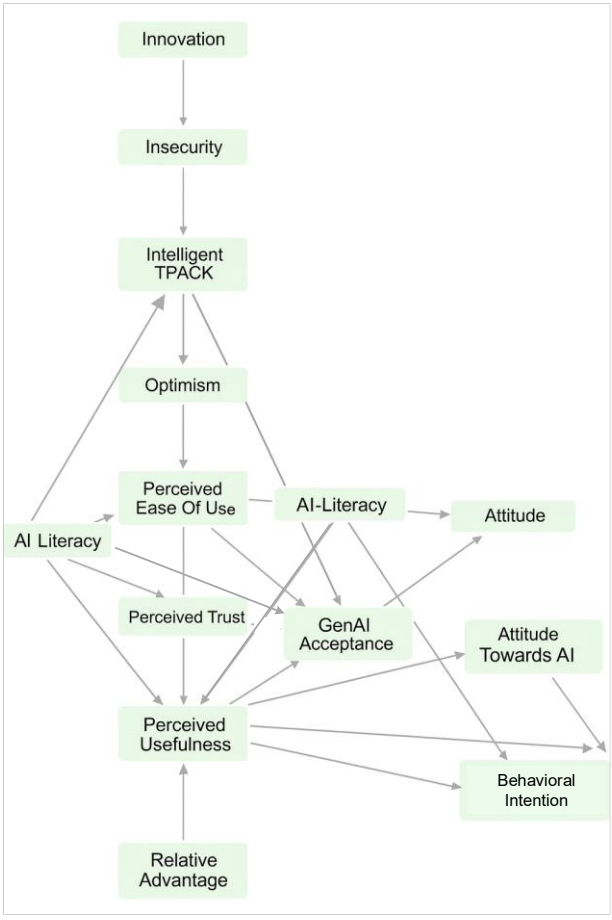


Fig. 6. AI adoption model.

1) RQ1: Readiness of Higher Education to Adopt AI and its Relevance to Continuing Education

Integrating AI into teaching through the Intelligent-TPACK lens presupposes pedagogical expertise, innovation readiness,

AI literacy, and robust teacher trust [44]. Because AI functions not merely as a tool but as a catalyst for pedagogical change, instructors must engage in reflective and ethical practice. In this view, the success of AI adoption hinges on educators' capacity to build Intelligent-TPACK. AI can also strengthen education by advancing civic responsibility, ethics, and character formation; realizing this potential requires AI-literate human resources and curricula that embed AI ethics [58]. To ensure students act as responsible, productive users rather than passive consumers, higher education institutions should elevate AI literacy—thereby supporting engagement and personalized learning. The upshot is that AI literacy is essential to cultivate ethical and critical AI use [59].

Within faculty recruitment, the use of GenAI renders AI literacy and Intelligent-TPACK foundational prerequisites. Framed this way, AI promotes data-driven, inclusive learning and supports the transition toward sustainable education. Institutions must therefore secure adequate digital infrastructure, ethical policies, and a supportive ecosystem—backed by AI training for lecturers [60].

SLR findings indicate heterogeneous readiness across systems. Universities in developed contexts (e.g., China, the United States, Australia) have integrated AI into adaptive learning, automated assessment, and academic chatbots, underwritten by mature infrastructure and regulation. These findings enable attention not only to adoption but also to sustainability dimensions such as energy efficiency, learning inclusion, and 21st-century literacies. By contrast, in many developing countries, readiness remains constrained by infrastructural gaps, limited funding, low digital literacy, and absent institutional policies, leaving the contribution of AI to sustainable education underrealized due to the digital divide. In Indonesia, despite government initiatives for educational digitalization, institutional readiness varies widely. Thus, AI readiness can serve as an accelerator of sustainable education—but only alongside deliberate strategies to build AI literacy, institute ethical governance, and secure national policy support.

2) RQ2: Potential for Cross-Theoretical Integration in the AI Adoption Conceptual Framework

UTAUT2 delineates determinants of behavioral intention for AI-based virtual assistants—performance expectancy, effort expectancy, social influence, and facilitating conditions—making it readily extensible to AI adoption in higher education. By contrast, TAM concentrates on perceived usefulness and related perception constructs while largely omitting social-infrastructure considerations. Treating these frameworks in isolation yields a piecemeal view of adoption; integrating TAM and UTAUT2 is therefore essential to capture AI's multidimensional uptake rather than privileging either technical or social facets alone [61]. Complementarily, TAM remains a primary predictor of technology acceptance, whereas TPB (Theory of Planned Behavior) enriches the socio-psychological layer. Together, TAM and TPB furnish a sustainable scaffold by combining technical, psychological, and social drivers, thereby strengthening inclusive science learning [62]. Because adoption in educational settings also hinges on pedagogy, TPACK serves as a gauge of teacher readiness; TPACK literacy tends to increase teachers' acceptance of AI's usefulness, particularly as roles shift from information transmitters to facilitators of data-driven, personalized learning [44].

Most studies still deploy TAM to assess usefulness and ease, or UTAUT to add performance expectations and social influence; institution-level work commonly relies on TOE, and macro-level analyses sometimes invoke Diffusion of Innovation (DOI). This siloed use of frameworks fragments AI-adoption analysis. A cross-theoretical pathway is to combine TAM/UTAUT (individual level), TOE (organizational level), and TPACK (pedagogical level) within a single architecture. For instance, TAM can explain students' perceived usefulness of AI, TOE can capture organizational readiness (infrastructure, policies, external support), and TPACK can fortify lecturers' pedagogical and ethical literacy. Integrated in this way, the model becomes a holistic, multi-level framework suited not only to explaining adoption but also to designing implementation strategies aligned with sustainability goals—efficiency, inclusion, 21st-century literacy, and digital ethics.

3) RQ3: Contribution of Developing Countries in Realizing AI Adoption-Based Sustainable Education

AI is a powerful lever for improving learning quality, equitable access, and administrative efficiency. In developing contexts, however, success turns critically on the readiness of human resources and infrastructure; persistent challenges include infrastructural deficits, policy gaps, ethical risks, and limited AI literacy. These constraints can become opportunities when AI is positioned as an accelerator of educational transformation through policy support, capacity building, and targeted technology investment [63]. AI-based chatbots can also augment decision-making by enabling real-time interaction, personalized responses, and rapid data analysis, thereby elevating service quality and advancing continuing education as a transformational innovation [64]. Evidence further indicates that AI tools can raise academic performance by increasing learning efficiency, though concerns about personal-data misuse and algorithmic transparency—and the attendant risks of bias—underscore that safety and ethics are integral to sustainable education, not optional add-ons [65]. In the Middle East, university sustainability ratings have improved alongside AI adoption; beyond academic efficiency, AI has helped attract international students and spurred curricular shifts toward digitalization, AI ethics, and the green economy—together reinforcing AI's direct contribution to sustainability ratings in the region [66].

Developing countries such as Indonesia, Malaysia, and Sri Lanka are beginning to contribute empirical work on AI literacy, automated assessment, and generative AI ethics. Their studies provide essential local context to complement East Asia's dominant literature—for example, funding and infrastructure constraints in Nigeria and the absence of university-level regulation in Sri Lanka. To advance AI-enabled sustainable education, developing countries can:

Present local context-based models—for example, integrating AI with distance learning in remote areas.

Contribute policy perspectives—examining how governments, universities, and communities can narrow the digital divide.

Develop ethics and AI-literacy research—recognizing that the central challenge is not only technological but also human capacity and governance.

Pursued together, these strategies position developing countries not merely as consumers of global scholarship but as producers of contextually grounded knowledge that broadens and deepens the meaning of continuing education.

4) RQ4: Relationship of AI Adoption Conceptual Framework to Sustainable Education Goals

AI adoption for continuing education is multi-variable in nature. Perceived Usefulness and Optimism are pivotal because uptake in higher education hinges on users' confidence in the technology [67], and positive attitudes among students, institutions, and lecturers further reinforce acceptance. Conversely, inadequate AI literacy and concerns over data-security risks act as barriers. These conditions imply that successful adoption requires a coordinated blend of infrastructure readiness, policy support, and human-resource development. In this sense, AI is not merely a tool but a driver of continuing education when embedded within an integrated digital-transformation strategy. Related evidence indicates that trust, privacy, and perceived usefulness are necessary to encourage use, while enjoyment is important for cultivating student engagement; adoption also turns on psychological factors, since safe and positive user experiences shape success [45]. Accordingly, effective AI integration must address not only technical efficiencies but also social, ethical, and psychological dimensions [61].

The SLR-synthesized framework positions individual (digital literacy, trust, motivation), institutional (management support, infrastructure, campus policy), and external (regulation, global trends, ethics) factors as primary determinants of AI adoption. Their interplay forms an adoption ecosystem that contributes directly to sustainable-education targets: 1) SDG 4 (Quality Education): AI advances personalization, adaptive assessment, and inclusive access; 2) SDG 9 (Industry, Innovation, and Infrastructure): universities operate as AI innovation hubs that strengthen digital infrastructure; 3) SDG 10 (Reduced Inequalities): AI enables inclusion for remote learners via smart distance education; 4) SDG 17 (Partnerships for the Goals): cross-sector collaboration among universities, government, and industry in AI development. Thus, the conceptual framework not only maps determinants of adoption but also traces a causal pathway through which AI becomes an instrument for sustainable education—especially salient for developing countries working to narrow the digital divide.

IV. Conclusion

This study systematically reviewed AI adoption in higher education and its links to sustainable education. The evidence shows that perceived usefulness, ease of use, trust, optimism, and AI literacy strongly shape adoption behavior. However, the literature remains fragmented—dominated by single-theory applications (e.g., TAM or UTAUT) and concentrated in developed regions. To address these gaps, we propose an integrated, multi-level framework that combines TAM, UTAUT, TPACK, and TOE to capture individual, institutional, and policy dimensions. The synthesis indicates that AI adoption enhances efficiency, personalization, and academic performance while directly advancing sustainability objectives, notably SDG 4 (Quality Education) and SDG 10 (Reduced Inequalities). Future research should prioritize longitudinal designs, ethics–policy integration, and context-sensitive models in developing countries. In sum, AI should be understood not only as a technological innovation but as a transformative lever for accelerating equitable and sustainable higher education.

V. Limitations and Future Research

This study acknowledges several limitations. First, the data were drawn exclusively from Scopus-indexed, English-language articles published between 2015 and 2025, and the chosen keyword strategy may have omitted relevant works that did not explicitly emphasize the term "adoption." Second, technological developments—particularly the rapid evolution of GenAI after 2022—progress so swiftly that some findings may lag behind emerging practices. Third, the heavy theoretical reliance on individual-level frameworks such as TAM and UTAUT, coupled with regional imbalances, limits generalizability. To address these gaps, future studies should 1) test an integrated, multilevel TAM–UTAUT–TPACK–TOE model using hierarchical modelling, and 2) connect adoption measures to objective indicators through longitudinal designs while broadening contextual analyses.

Acknowledgment

This research was supported by BIMA Kemendiktisaintek (Indonesian Ministry of Higher Education, Science, and Technology) under the Fundamental Research (PFR) Scheme, Fiscal Year 2025, through the main contract No. 091/C3/DT.05.00/PL/2025 and the derivative contract No. B/17.41/UN34.9/PT/2025

References

[1] K.-E.-K. Babu, "Artificial Intelligence, Its Applications in Different Sectors and Challenges: Bangladesh Context," 2021, pp. 103-119. doi: 10.1007/978-3-030-88040-8_4.

[2] D. Schiff, "Education for AI, not AI for Education: The Role of Education and Ethics in National AI Policy Strategies," Int J Artif Intell Educ, vol. 32, no. 3, pp. 527-563, Sep. 2022, doi: 10.1007/s40593-021-00270-2.

[3] B. George and O. Wooden, "Managing the Strategic Transformation of Higher Education through Artificial Intelligence," Adm Sci, vol. 13, no. 9, p. 196, Aug. 2023, doi: 10.3390/admsci13090196.

[4] N. Subchan, "Conceptual Framework of Innovative Library Services Based on Artificial Intelligence (AI) in Order to Accelerate Digital Transformation," JPUA: Airlangga University Library Journal: Librarianship Information & Communication Media, vol. 14, no. 1, pp. 1-14, 2024.

[5] D. Aggarwal, D. Sharma, and A. B. Saxena, "Exploring the Role of Artificial Intelligence for Augmentation of Adaptable Sustainable Education," Asian Journal of Advanced Research and Reports, vol. 17, no. 11, pp. 179-184, Oct. 2023, doi: 10.9734/ajarr/2023/v17i11563.

[6] H. S. Rad, R. Alipour, and A. Jafarpour, "Using Artificial Intelligence to Foster Students' Writing Feedback Literacy, Engagement, and Outcome: A Case of Wordtune Application," Interactive Learning Environments, vol. 32, no. 9, pp. 5020-5040, Oct. 2024, doi: 10.1080/10494820.2023.2208170.

[7] T. Thomas, "The Role of Artificial Intelligence in Formal and Infomal Education for Students," Int J Res Appl Sci Eng Technol, vol. 12, no. 3, pp. 69-71, 2024.

[8] S. Omidvar and H. Meihami, "Exploring the 'What' and 'How' of Opportunities and Challenges of AI in EFL Teacher Education,"

Computers and Education: Artificial Intelligence, vol. 9, p. 100443, Dec. 2025, doi: 10.1016/j.caeai.2025.100443.

[9] R. Tiwari, "The Integration of AI and Machine Learning in Education and its Potential to Personalize and Improve Student Learning Experiences," International Journal of Scientific Research in Engineering and Management (IJSREM), vol. 7, no. 2, pp. 1-11, 2023.

[10] M. Mehrvarz, G. Salimi, S. Abdoli, and B. M. McLaren, "How does Students' Perception of ChatGPT Shape Online Learning Engagement and Performance?", Computers and Education: Artificial Intelligence, vol. 9, pp. 1-15, Dec. 2025, doi: 10.1016/j.caeai.2025.100459.

[11] J. Iqbal, V. Asgarova, Z. F. Hashmi, B. N. Ngajie, M. Z. Asghar, and H. Järvenoja, "Exploring Faculty Experiences with Generative Artificial Intelligence Tools Integration in Second Language Curricula in Chinese Higher Education," Discover Computing, vol. 28, no. 1, pp. 1-26, Jun. 2025, doi: 10.1007/s10791-025-09655-6.

[12] O. Zawacki-Richter and I. Jung, "Handbook of Open, Distance and Digital Education," Springer, 2023.

[13] J. Zhang, "Impact of Artificial Intelligence on Higher Education in the Perspective of its Application of Transformation," Lecture Notes in Education Psychology and Public Media, vol. 2, no. 1, pp. 822-830, 2023.

[14] H. Crompton and D. Song, "The Potential of Artificial Intelligence in Higher Education," Revista Virtual Universidad Católica del Norte, no. 62, pp. 1-4, Jan. 2021, doi: 10.35575/rvucn.n62a1.

[15] M. P. Ilić, D. Păun, N. Popović Šević, A. Hadžić, and A. Jianu, "Needs and Performance Analysis for Changes in Higher Education and Implementation of Artificial Intelligence, Machine Learning, and Extended Reality," Educ Sci (Basel), vol. 11, no. 10, p. 568, Sep. 2021, doi: 10.3390/educsci11100568.

[16] R. Sandu, E. Gide, and M. Elkhodr, "The Role and Impact of ChatGPT in Educational Practices: Insights from an Australian Higher Education Case Study," Discover Education, vol. 3, no. 1, p. 71, Jun. 2024, doi: 10.1007/s44217-024-00126-6.

[17] S. F. Ahmad, Mohd. K. Rahmat, M. S. Mubarik, M. M. Alam, and S. I. Hyder, "Artificial Intelligence and Its Role in Education," Sustainability, vol. 13, no. 22, p. 12902, Nov. 2021, doi: 10.3390/su132212902.

[18] Z. Xu, "AI in education: Enhancing learning experiences and student outcomes," Applied and Computational Engineering, vol. 51, no. 1, pp. 104-111, Mar. 2024, doi: 10.54254/2755-2721/51/20241187.

[19] L. Hughes, T. Malik, S. Dettmer, A. S. Al-Busaidi, and Y. K. Dwivedi, "Reimagining Higher Education: Navigating the Challenges of Generative AI Adoption," Information Systems Frontiers, Feb. 2025, doi: 10.1007/s10796-025-10582-6.

[20] O. Festus and O. Bamidele Emmanuel, "Sociocultural and Digital Communication Challenges in AI Adoption for Classroom Communication: Insights from Nigerian Colleges of Education," Language, Technology, and Social Media, Dec. 2024, doi: 10.70211/ltsm.v3i1.115.

[21] X. Wang, S. Zhao, X. Xu, H. Zhang, and V. N.-L. Lei, "AI adoption in Chinese universities: Insights, challenges, and opportunities from academic leaders," Acta Psychol (Amst), vol. 258, p. 105160, Aug. 2025, doi: 10.1016/j.actpsy.2025.105160.

[22] A. M. Al-Zahrani and T. M. Alasmari, "A Comprehensive Analysis of AI Adoption, Implementation Strategies, and Challenges in Higher Education Across the Middle East and North Africa (Mena) Region," Educ Inf Technol (Dordr), vol. 30, no. 8, pp. 11339-11389, Jun. 2025, doi: 10.1007/s10639-024-13300-y.

[23] A. Henadirage and N. Gunarathne, "Barriers to and Opportunities for the Adoption of Generative Artificial Intelligence in Higher Education in the Global South: Insights from Sri Lanka," Int J Artif Intell Educ, vol. 35, no. 1, pp. 245-281, Mar. 2025, doi: 10.1007/s40593-024-00439-5.

[24] A. Behl, M. Chavan, K. Jain, I. Sharma, V. E. Pereira, and J. Z. Zhang, "The Role of Organizational Culture and Voluntariness in the Adoption of Artificial Intelligence for Disaster Relief Operations," Int J Manpow, vol. 43, no. 2, pp. 569-586, May 2022, doi: 10.1108/IJM-03-2021-0178.

[25] Y. Chen, Y. Hu, S. Zhou, and S. Yang, "Investigating the Determinants of Performance of Artificial Intelligence Adoption in Hospitality Industry During COVID-19," International Journal of Contemporary Hospitality Management, vol. 35, no. 8, pp. 2868-2889, Jul. 2023, doi: 10.1108/IJCHM-04-2022-0433.

[26] Y. Kim, V. Blazquez, and T. Oh, "Determinants of Generative AI System Adoption and Usage Behavior in Korean Companies: Applying the UTAUT Model," Behavioral Sciences, vol. 14, no. 11, p. 1035, Nov. 2024, doi: 10.3390/bs14111035.

[27] M. A. Nugroho, "Impact of Government Support and Competitor Pressure on the Readiness of SMEs in Indonesia in Adopting the Information Technology," Procedia Comput Sci, vol. 72, pp. 102-111, 2015, doi: 10.1016/j.procs.2015.12.110.

[28] M. A. Nugroho, P. W. Dewanti, and B. T. Novitasari, "The Impact of Perceived Usefulness and Perceived Ease of Use on Student's Performance in Mandatory E-Learning Use," in 2018 International Conference on Applied Information Technology and Innovation (ICAITI), IEEE, Sep. 2018, pp. 26-30. doi: 10.1109/ICAITI.2018.8686742.

[29] M. A. Nugroho and M. A. Fajar, "Effects of Technology Readiness Towards Acceptance of Mandatory Web-Based Attendance System," Procedia Comput Sci, vol. 124, pp. 319-328, 2017, doi: 10.1016/j.procs.2017.12.161.

[30] M. A. Nugroho, R. Jusoh, and N. A. M. Salleh, "The Role of Alignment Between IS Strategy and Social Capital on the IS Capability and Business Performance Relationship: A Cross-Sectional Survey," IEEE Access, vol. 8, pp. 152760-152771, 2020, doi: 10.1109/ACCESS.2020.3018036.

[31] M. A. Nugroho and B. T. Novitasari, "Fintech Risks and Continuance to Use on Generation Z," Journal of Law and Sustainable Development, vol. 11, no. 2, p. e630, Jul. 2023, doi: 10.55908/sdgs.v11i2.630.

[32] M. A. Nugroho, B. T. Novitasari, and R. P. Timur, "The Mediating Role of E-Commerce Adoption in the Relationship Between Government Support and SME Performance in Developing Countries," International Journal of Sustainable Development and Planning, vol. 19, no. 3, pp. 1023-1032, Mar. 2024, doi: 10.18280/ijsdp.190320.

[33] M. A. Nugroho, D. Rahmawati, and B. T. Novitasari, "The Influence of Website Quality on E-Learning Usage Continuity," Journal of Advanced Research in Dynamical and Control Systems, vol. 11, no. 11, pp. 382-388, 2019.

[34] M. A. Nugroho, D. Setyorini, and B. T. Novitasari, "The Role of Satisfaction on Perceived Value and E-Learning Usage Continuity Relationship," Procedia Comput Sci, vol. 161, pp. 82-89, 2019, doi: 10.1016/j.procs.2019.11.102.

[35] Mu'ah, U. Y. Suyanto, K. D. Syaputro, S. Musarofah, and N. Qomariah, "Increasing Customers Loyalty MSME of Focused E-Marketing and Quality of Service," International Journal of Engineering Research and Technology, vol. 13, no. 10, p. 2729, Oct. 2020, doi: 10.37624/IJERT/13.10.2020.2729-2735.

[36] M. B. Triyono, A. A. Rafiq, D. Hariyanto, D. Adinda, and M. Denami, "In-World NPC: Analyzing Artificial Intelligence Precision in Virtual Reality Settings," International Journal of Online and Biomedical Engineering (iJOE), vol. 20, no. 15, pp. 19-42, Dec. 2024, doi: 10.3991/ijoe.v20i15.51437.

[37] A. Alrayes, T. F. Henari, and D. A. Ahmed, "ChatGPT in Education - Understanding the Bahraini Academics Perspective," Electronic Journal of e-Learning, vol. 22, no. 2, pp. 112-134, Apr. 2024, doi: 10.34190/ejel.22.2.3250.

[38] J.-I. Choi, E. Yang, and E.-H. Goo, "The Effects of an Ethics Education Program on Artificial Intelligence among Middle School Students: Analysis of Perception and Attitude Changes," Applied Sciences, vol. 14, no. 4, p. 1588, Feb. 2024, doi: 10.3390/app14041588.

[39] D. W. Purnama, C. B. A. Rochman, R. Darmawan, P. Amaliyah, A. D. Baqi, and A. Zahidin, "A Bibliometrics Analysis for Artificial Intelligence Implementation of Employment in Education Institutions," INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH AND ANALYSIS, vol. 07, no. 01, pp. 7-16, Jan. 2023, doi: 10.47191/ijmra/v7-i01-02.

[40] E. S. Tenakwah et al., "Generative AI and Higher Education Assessments: A Competency-Based Analysis," Aug. 03, 2023. doi: 10.21203/rs.3.rs-2968456/v2.

[41] I. H. Y. Yim and R. Wegerif, "Teachers' Perceptions, Attitudes, and Acceptance of Artificial Intelligence (AI) Educational Learning Tools: An Exploratory Study on AI Literacy for Young Students," Future in

Educational Research, vol. 2, no. 4, pp. 318-345, Dec. 2024, doi: 10.1002/fer3.65.

[42] N. I. Mohd Rahim, N. A. Iahad, A. F. Yusof, and M. A. Al-Sharafi, "AI-Based Chatbots Adoption Model for Higher-Education Institutions: A Hybrid PLS-SEM-Neural Network Modeling Approach," sustainability, vol. 14, no. 19, p. 12726, Oct. 2022, doi: 10.3390/su141912726.

[43] I. Y. Alyoussef, A. M. Drwish, F. A. Albakheet, and R. H. Alhajhoj, "AI Adoption for Collaboration: Factors Influencing Inclusive Learning Adoption in Higher Education," IEEE Access, vol. 13, pp. 81690-81713, 2025.

[44] I. Celik, "Towards Intelligent-TPACK: An Empirical Study on Teachers' Professional Knowledge to Ethically Integrate Artificial Intelligence (AI)-Based Tools into Education," Comput Human Behav, vol. 138, p. 107468, Jan. 2023, doi: 10.1016/j.chb.2022.107468.

[45] G. McLean and K. Osei-Frimpong, "Hey Alexa ... Examine the Variables Influencing the Use of Artificial Intelligent In-Home Voice Assistants," Comput Human Behav, vol. 99, pp. 28-37, Oct. 2019, doi: 10.1016/j.chb.2019.05.009.

[46] S. K. Rao and R. Prasad, "Impact of 5G Technologies on Industry 4.0," Wirel Pers Commun, vol. 100, no. 1, pp. 145-159, May 2018, doi: 10.1007/s11277-018-5615-7.

[47] R. O. Okunlaya, N. Syed Abdullah, and R. A. Alias, "Artificial Intelligence (AI) Library Services Innovative Conceptual Framework for the Digital Transformation of University Education," Library Hi Tech, vol. 40, no. 6, pp. 1869-1892, Dec. 2022, doi: 10.1108/LHT-07-2021-0242.

[48] C.-C. Lin, A. Y. Q. Huang, and O. H. T. Lu, "Artificial Intelligence in Intelligent Tutoring Systems Toward Sustainable Education: a Systematic Review," Smart Learning Environments, vol. 10, no. 1, p. 41, Aug. 2023, doi: 10.1186/s40561-023-00260-y.

[49] F. Kamalov, D. Santandreu Calonge, and I. Gurrib, "New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution," sustainability, vol. 15, no. 16, p. 12451, Aug. 2023, doi: 10.3390/su151612451.

[50] S. Xu, K. F. Kee, W. Li, M. Yamamoto, and R. E. Riggs, "Examining the Diffusion of Innovations from a Dynamic, Differential-Effects Perspective: A Longitudinal Study on AI Adoption Among Employees," Communic Res, vol. 51, no. 7, pp. 843-866, 2024.

[51] D. Tverskoi, S. Babu, and S. Gavrilets, "The Spread of Technological Innovations: Effects of Psychology, Culture and Policy Interventions," R Soc Open Sci, vol. 9, no. 6, p. 211833, Jun. 2022, doi: 10.1098/rsos.211833.

[52] J.-H. Han and H. J. Sa, "Acceptance of and Satisfaction with Online Educational Classes through the Technology Acceptance Model (TAM): The COVID-19 Situation in Korea," Asia Pacific Education Review, vol. 23, no. 3, pp. 403-415, Sep. 2022, doi: 10.1007/s12564-021-09716-7.

[53] G. Lame, "Systematic literature reviews: An introduction," in Proceedings of the design society: international conference on engineering design, Cambridge University Press, 2019, pp. 1633-1642.

[54] A. Nightingale, "A Guide to Systematic Literature Reviews," Surgery (Oxford), vol. 27, no. 9, pp. 381-384, Sep. 2009, doi: 10.1016/j.mpsur.2009.07.005.

[55] S. Kraus, M. Breier, and S. Dasí-Rodríguez, "The Art of Crafting a Systematic Literature Review in Entrepreneurship Research," International Entrepreneurship and Management Journal, vol. 16, no. 3, pp. 1023-1042, Sep. 2020, doi: 10.1007/s11365-020-00635-4.

[56] M. L. Rethlefsen et al., "PRISMA-S: An Extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews," Syst Rev, vol. 10, no. 1, p. 39, Jan. 2021, doi: 10.1186/s13643-020-01542-z.

[57] L. Shamseer et al., "Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P) 2015: Elaboration and Explanation," BMJ, vol. 349, no. jan02 1, pp. g7647-g7647, Jan. 2015, doi: 10.1136/bmj.g7647.

[58] T. Dong, T. He, and F. Pang, "The Path of Civic Education and Intelligent Strategies for Counselors in the Background of Artificial Intelligence," Applied Mathematics and Nonlinear Sciences, vol. 9, no. 1, Jan. 2024, doi: 10.2478/amns.2023.2.00821.

[59] A. M. Al-Abdullatif and M. A. Alsubaie, "ChatGPT in Learning: Assessing Students' Use Intentions through the Lens of Perceived Value and the Influence of AI Literacy," Behavioral Sciences, vol. 14, no. 9, p. 845, Sep. 2024, doi: 10.3390/bs14090845.

[60] A. M. Al-Abdullatif, "Modeling Teachers' Acceptance of Generative Artificial Intelligence Use in Higher Education: The Role of AI Literacy, Intelligent TPACK, and Perceived Trust," Educ Sci (Basel), vol. 14, no. 11, p. 1209, Nov. 2024, doi: 10.3390/educsci14111209.

[61] M. García de Blanes Sebastián, J. R. Sarmiento Guede, and A. Antonovica, "Application and Extension of the UTAUT2 Model for Determining Behavioral Intention Factors in Use of Artificial Intelligence Virtual Assistants," Front Psychol, vol. 13, Oct. 2022, doi: 10.3389/fpsyg.2022.993935.

[62] O. P. Adelana, M. A. Ayanwale, and I. T. Sanusi, "Exploring Pre-Service Biology Teachers' Intention to Teach Genetics Using an AI Intelligent Tutoring-Based System," Cogent Education, vol. 11, no. 1, p. 2310976, Dec. 2024, doi: 10.1080/2331186X.2024.2310976.

[63] O. Hamal, N.-E. El Faddouli, M. H. A. Harouni, and J. Lu, "Artificial intelligence in education," sustainability, vol. 14, no. 5, p. 2862, Mar. 2022, doi: 10.3390/su14052862.

[64] J. N. K. Wah, "Revolutionizing E-Health: the Transformative Role of AI-Powered Hybrid Chatbots in Healthcare Solutions," Front Public Health, vol. 13, pp. 1-24, Feb. 2025, doi: 10.3389/fpubh.2025.1530799.

[65] J. T. K. Phua, H.-F. Neo, and C.-C. Teo, "Evaluating the Impact of Artificial Intelligence Tools on Enhancing Student Academic Performance: Efficacy Amidst Security and Privacy Concerns," Big Data and Cognitive Computing, vol. 9, no. 5, p. 131, May 2025, doi: 10.3390/bdcc9050131.

[66] M. H. Ronaghi and M. Ronaghi, "How Does the Use of Artificial Intelligence Affect Sustainability Rating in Middle Eastern Universities?," Asian Education and Development Studies, vol. 14, no. 2, pp. 152-172, Mar. 2025, doi: 10.1108/AEDS-08-2024-0182.

[67] W. Hadid, S. Horii, and A. Yokota, "Artificial Intelligent Technologies in Japanese Manufacturing Firms: An Empirical Survey Study," Int J Prod Res, vol. 63, no. 1, pp. 193-219, Jan. 2025, doi: 10.1080/00207543.2024.2358409.

Comparative Review of Confidence and Other Evaluation Metrics in Predictive Modeling for Procurement Fraud Coalition

Saifuddin Mohd, Mohamad Taha Ijab
Institute of Visual Informatics, Universiti Kebangsaan Malaysia, 43600 Bangi Selangor, Malaysia

***Abstract*—Procurement fraud, particularly when bidders act together through collusion or coalition schemes, remains a major threat to fair competition in public procurement. Predictive modeling has emerged as a key analytical tool for detecting such behaviors yet choosing appropriate evaluation metrics continues to be a challenge, especially with imbalanced or correlated data. This study applies a structured narrative review supported by a comparative analysis to examine commonly used evaluation metrics—Accuracy, Precision, Recall, F1-score, and AUC-ROC—in relation to the rule-based Confidence metric derived from association rule mining. The findings reveal that while traditional classification metrics are effective for general predictive tasks, they often fail to capture relational and co-occurrence patterns that characterize coalition fraud. In contrast, Confidence demonstrates higher interpretability and contextual relevance for detecting collusive behaviors among suppliers. The study highlights the potential of hybrid evaluation frameworks that combine classification and rule-based measures to improve fraud detection accuracy and explainability. This approach contributes to advancing predictive modeling, procurement analytics, and coalition detection by emphasizing metrics that balance performance, interpretability, and real-world applicability.**

***Keywords*—Procurement fraud; predictive modeling; confidence; evaluation metrics; association rule mining; coalition detection; public sector analytics**

I. Introduction

Public administration rests on a simple promise to citizens: public money must be spent properly and fairly. Public procurement sits at the heart of that promise, shaping how governments acquire goods and services at scale. Digitisation has helped by opening processes to scrutiny and producing rich data that can be analysed for fairness and efficiency. At the same time, the same openness has introduced new risks as actors learn to exploit procedural signals and digital traces for gain [1], [2].

Among those risks, collusion fraud is especially damaging. It arises when suppliers coordinate to shape outcomes through bid rigging, rotating winners, or submitting coordinated bids that blunt competition. These schemes are difficult to spot because the telltale signs rarely appear in a single tender. They live in patterns across time and across networks of bidders, which standard document checks or routine audits often miss [3].

Analytics has therefore become a practical route to detection. Predictive models sift large datasets to flag anomalies that suggest misconduct [4] The challenge is how we judge these models in a setting where fraud is rare. Accuracy, precision, recall, F1-score and AUC-ROC are widely used, yet each can mislead in imbalanced data. A model can be "accurate" by predicting almost everything as clean. These metrics also focus on isolated transactions and often ignore the relational signals that define collusion, such as repeated co-bidding or synchronized entry and withdrawal patterns [4], [5].

This study brings a complementary lens to the problem by examining Confidence from association rule mining as an evaluation and insight metric. Confidence estimates the likelihood of an event given another event. In procurement terms, it answers questions like "how often does Supplier B bid when Supplier A bids." That simple conditional view helps surface co-occurrence patterns that are consistent with coordinated behaviour and gives analysts something they can read and explain [6].

Although predictive analytics has matured, how we evaluate models for collusion specifically remains underexplored. Collusion is relational by nature. It emerges from ties, routines, and repeated interactions among bidders that traditional classification metrics do not capture well [7], [8], [9]. Ignoring those ties risks high-scoring models that miss the very behaviours of interest.

Therefore, this paper compares Confidence against common performance metrics and argue for hybrid evaluation. Classification metrics remain useful for measuring statistical performance. Confidence and related rule-based measures supply contextual signals about who appears with whom, and how often, in ways auditors and regulators can act on. The combination improves detection quality and interpretability in practice [10], [11]. This integrated focus aligns with the intent of the present manuscript, which positions Confidence alongside accuracy, precision, recall, F1-score and AUC-ROC for coalition detection in public procurement, and motivates a balanced, insight-driven assessment approach.

The rest of the paper proceeds as follows. Section II reviews literature on collusion in procurement and on evaluation metrics for predictive modelling, with emphasis on Confidence. Section III explains the methodology. Section IV discusses the comparative analysis and findings. Section V presents the synthesis and discussion and Section VI concludes and outlines directions for future research in predictive procurement analytics.

II. Background and Related Work

A. Coalition Fraud in Public Procurement

It is quite easy to deceive and change the way the government buys things. Coalition fraud is one of the most advanced and harmful types of fraud. Collusion, bid rigging, or cartel activity occurs when two or more suppliers work together to change the outcome of a procurement process. These kinds of alliances are meant to get rid of real competition. They accomplish this by bidding together over and over again, swapping the winner before the auction ends, or making bids that are very similar to each other.

These kinds of alliances are meant to get rid of real competition. They accomplish this by bidding together over and over again, swapping the winner before the auction ends, or making bids that are very similar to each other. These actions not only mess up how the market works, but they also make people lose a lot of trust and make items too expensive to acquire [12], [13]. Coalition fraud is always about connections, unlike single fraudulent activities. It takes advantage of how hard it is to grasp the different parts of the evaluation and how often bidding groups act the same way. This makes it challenging to find occurrences using typical techniques of assessment.

The Malaysia Competition Commission (MyCC) has observed that in Malaysia's public e-procurement settings, behavior that appears like a cartel is growing more sophisticated. We need to find it using improved, data-driven ways [14].

B. Evaluation Metrics in Predictive Modeling

In predictive modeling for procurement fraud detection, assessment metrics are necessary to measure both accuracy and efficacy. The most fundamental indicator is accuracy, which shows how many of the predictions were accurate. However, in datasets with significant imbalance, frequently seen in fraud detection, accuracy can be deceptive as it may overstate performance by privileging the majority class [15], [16].

Precision and recall are two indicators that provide you more information. Precision measures how many of the highlighted instances were indeed frauds. This is very important in situations when false positives might create reputational or administrative costs. On the other hand, recall evaluates how well a model can find all real fraud cases, which is very important in high-stakes situations where missing fraud can have serious consequences [17], [18]. The F1-score is a harmonic mean of accuracy and recall, which is typically used to give a fair assessment because these metrics often disagree with each other [19]. In procurement fraud detection, however, recall is frequently more important than precision since missing fraud is more damaging than overestimating its presence.

Other metrics, such as AUC-ROC, are useful since they do not depend on a certain threshold and can tell the difference between different things [20]. But since they are so abstract, they are not very good for rule-based or coalition-oriented fraud detection, where patterns that are relevant to the situation are important. Recent advancements in Explainable AI (XAI) highlight the necessity for metrics that provide transparency and interpretability, ensuring that model outputs can be successfully conveyed to policymakers, auditors, and non-technical stakeholders [21].

C. Confidence as a Metric for Coalition Detection

Given the limitations of global evaluation metrics, rule-based measures, particularly confidence from association rule mining, have garnered increasing attention. Confidence measures the conditional probability of a consequent given an antecedent, formally expressed as $P(Y|X)$. In the context of procurement fraud, this can be interpreted as the likelihood of a bidder participating or winning given the participation of another supplier, thus offering critical insight into co-occurrence behavior and potential [22], [23]. Unlike aggregate classification metrics, confidence allows for localized, pattern-specific interpretability.

This makes it particularly effective for identifying relational anomalies, such as recurrent supplier pairings or tenders with suspiciously homogeneous bidder compositions. Furthermore, confidence can be used in unsupervised settings, making it applicable even when labeled fraud data is scarce and often-cited challenge in procurement analytics. Nonetheless, confidence is not without limitations. It can be artificially inflated by low-support rules, leading to overinterpretation of spurious associations. To address this, confidence is typically used in conjunction with support (frequency of rule occurrence) and lift (statistical strength relative to random chance), forming a robust triad of rule evaluation metrics [23].

III. Methodology

In examining studies on procurement fraud and coalition detection, two methodologies for synthesizing information are commonly utilized: structured narrative reviews and systematic reviews. Each approach has various advantages and disadvantages. Systematic reviews are recognized for their methodological rigor, enhancing transparency, reproducibility, and confidence in the resultant evidence [24]. The rigor is especially crucial when looking at complex evaluation criteria for predictive modeling in procurement fraud, where the accuracy of the results is critical [25]. The strict guidelines of systematic reviews may make it harder to look into new ideas or processes, which are sometimes better handled using more flexible methods [26].

In Contrast, structured narrative reviews, on the allow researchers combine results from many different sources of literature without having to follow strict rules on the methodological guidelines [27]. This adaptability supports the formation of nuanced findings, particularly in coalition detection, where conceptual, empirical, and methodological perspectives intersect [28]. A structured narrative review is not only descriptive; it seeks to clarify linkages across contemporary studies, uncover methodological strengths, and pinpoint gaps for further study.

This study employs a structured narrative review method to examine the Confidence metric in predictive modeling for procurement fraud and compare it to traditional classification-based evaluation metrics. This technique emphasizes the study's focus on relational and rare coalition activities, requiring a balance between comprehensive synthesis and interpretative depth. The review is conducted with inclusion and exclusion

criteria, a clearly defined topic, and careful consideration of the narrative integration of results to ensure quality.

To enhance methodological transparency and replicability, a clear and systematic search strategy was developed. The review process followed four sequential phases:

- Identification: Relevant literature was identified through a comprehensive search across Scopus, Web of Science, and Google Scholar databases using defined keyword combinations. The search period covered 2010 to early 2025 to capture both foundational and emerging studies.
- Screening: Duplicate records were removed, and titles and abstracts were screened to ensure alignment with the research focus on evaluation metrics and coalition detection in procurement fraud.
- Eligibility: Full-text assessment was conducted to evaluate whether the articles met the inclusion criteria—namely, studies that (i) examined evaluation metrics used in fraud or anomaly detection, (ii) applied Confidence or association rule mining in predictive modeling, or (iii) focused on procurement, collusion, or cartel behavior detection.
- Quality Assessment: To maintain rigor, studies were appraised based on methodological soundness, clarity of metric application, and relevance to predictive modeling. Papers lacking sufficient methodological detail or relying solely on opinion-based commentary were excluded.

Exclusion criteria encompassed grey literature (e.g., blogs, white papers, or non-peer-reviewed content), studies unrelated to predictive modeling or procurement fraud, and those without measurable or interpretable performance metrics. This structured process ensured that the final body of literature was both methodologically robust and conceptually relevant. The resulting dataset of studies formed the empirical foundation for the comparative synthesis presented in later sections.

A. Research Design

The study is conceptualized as a comparative literature-based review supported by a case-informed synthesis. The objective is to evaluate the interpretability, contextual relevance, and sensitivity of common evaluation metrics: Accuracy, Precision, Recall, F1-score, and AUC-ROC against the rule-based Confidence metric. These metrics were selected due to their widespread use in fraud detection and classification modeling.

Given the exploratory nature of the study, a qualitative interpretive comparison is adopted rather than a purely statistical meta-analysis. This approach allows for incorporating insights from multiple domains (e.g., data mining, fraud analytics, public procurement) and ensures the findings remain grounded in operational and domain-specific requirements.

Limitations: It is important to acknowledge that the qualitative, narrative synthesis approach adopted in this study may limit quantitative generalizability. Findings presented here are primarily conceptual, based on expert interpretation and literature review rather than empirical validation. Therefore, future studies should consider conducting empirical analyses with large, real-world datasets to statistically validate the conceptual insights provided by this comparative assessment.

B. Data Sources and Literature Selection

The literature search was conducted across three major academic databases that is Scopus, Web of Science, and Google Scholar to ensure comprehensive coverage of both theoretical and applied studies. The search period spanned 2010 to early 2025, capturing the evolution of predictive modeling and fraud detection research.

The following keyword combinations were used to ensure search precision and breadth:

- "confidence metric" AND "fraud detection"
- "evaluation metrics" AND "predictive modeling"
- "procurement fraud" OR "coalition detection"
- "association rule mining" AND "bid rigging"

Each retrieved article underwent multi-stage screening as described in the methodology overview. Only peer-reviewed journal articles, conference papers, and technical reports were included, ensuring academic credibility.

The inclusion criteria emphasized empirical or conceptual contributions to predictive modeling, Confidence metric application, or procurement fraud detection. Studies focusing solely on algorithmic performance without interpretive or contextual analysis were excluded.

C. Analytical Framework

This study constructed a comparative framework based on three dimensions and systematically compare the metrics:

- Interpretability – the extent to which metric outputs can be meaningfully interpreted by human analysts and decision-makers;
- Sensitivity to Class Imbalance – the ability of the metric to provide reliable insights in datasets where fraud cases constitute a minority.
- Applicability to Coalition Detection – the metric's utility in modeling relational or co-occurrence patterns typical of procurement collusion.

Each metric was scored heuristically on a five-point Likert-type scale (1 = Low, 5 = High) for each dimension, drawing from synthesis of the literature and informed expert judgment. While such scoring does not provide definitive empirical quantification, it offers a structured and transparent means of articulating relative strengths and weaknesses among metrics. A common approach in methodological reviews is where direct comparisons are limited or context specific.

IV. Comparative Analysis and Findings

This section presents a comparative assessment of commonly used evaluation metrics in predictive fraud modeling, with particular emphasis on their relevance to coalition detection in procurement contexts. The metrics are examined in terms of their mathematical foundations, interpretability, sensitivity to

class imbalance, and contextual utility in identifying collusive behavior.

A. Metric Characteristics and Theoretical Basis

The principal characteristics of six commonly employed evaluation metrics: Accuracy, Precision, Recall, F1-score, AUC-ROC, and Confidence are summarised in Table I. While the first five metrics are grounded in classification theory and typically applied in supervised learning settings, Confidence originates from association rule mining and is often used in unsupervised or pattern mining contexts.

TABLE I. COMPARISON ACROSS SIX METRICS: ACCURACY, PRECISION, RECALL, F1-SCORE, AUC-ROC, AND CONFIDENCE

Metric	Mathematical Definition	Best Context	Advantages	Limitations
Accuracy	(TP + TN) / Total	Balanced datasets	Simple, standard	Biased with class imbalance
Precision	TP / (TP + FP)	High false positive cost	Positive outcome focus	Ignores false negatives
Recall	TP / (TP + FN)	High false negative cost	Captures all true positives	Ignores false positives
F1-Score	2 * (P * R) / (P + R)	Imbalanced data	Balances P and R	Hard to interpret directly
AUC-ROC	Area under the curve of TPR vs FPR	Model comparison, threshold-independent evaluation	Comprehensive discrimination measure	Lacks local interpretability
Confidence	P(Y|X)	Rule-based learning, fraud detection	Conditional, interpretable	Ignores minimum support

Accuracy, defined as the ratio of correct predictions (both positive and negative) to the total number of instances, is intuitive but problematic in imbalanced datasets. It often overstates performance by favoring the majority class, which in fraud detection typically represents non-fraudulent cases [29].

Precision and Recall offer more focused evaluations. Precision emphasizes the correctness of positive predictions (i.e., how many predicted frauds are true frauds), whereas Recall reflects the model's ability to capture all actual frauds. F1-score, the harmonic mean of Precision and Recall, provides a balanced measure, though it may lack intuitive clarity for stakeholders not versed in statistical evaluation [17].

AUC-ROC evaluates the classifier's ability to discriminate between classes across different threshold values. While it is threshold-independent, AUC lacks granularity and may obscure localized fraud patterns [20].

Confidence, expressed as P(Y|X), quantifies the conditional likelihood that event Y occurs given X. In procurement contexts, this could translate to the probability that one supplier bids or wins conditional on the participation of another. It thus enables fine-grained relational analysis and is particularly useful for uncovering suspicious co-bidding behavior that might indicate tacit collusion [22].

B. Comparative Matrix and Interpretive Scoring

A comparative assessment of six evaluation metrics commonly employed in procurement fraud detection, analyzed across two critical dimensions: interpretability, defined as the extent to which a metric can be intuitively understood and meaningfully applied by human decision makers, and imbalance sensitivity, referring to the metric's responsiveness to skewed class distributions, an inherent characteristic of fraud related datasets.

The results in Fig. 1 indicate that Confidence achieves the highest level of interpretability (score of 5) while exhibiting the lowest sensitivity to class imbalance (score of 1), positioning it as particularly advantageous for contexts where transparency and data imbalance are pressing concerns.

Conversely, accuracy demonstrates poor performance on both fronts, with low interpretability, especially at finer analytical levels, and a high susceptibility to data imbalance, thereby limiting its effectiveness in real world fraud detection scenarios. The scoring, which ranges from 1 (least favorable) to 5 (most favorable), is derived through synthesis of current literature and practitioner experience, an approach often adopted in review-based studies where direct empirical comparisons are scarce.

Confidence is rarely compared with traditional classification metrics such as Accuracy in empirical evaluations, necessitating a relative rather than absolute analytical approach. The purpose of this comparative framework is not to produce definitive rankings but to elucidate general tendencies in metric behavior, thereby informing the selection of evaluation strategies that align with the unique demands of procurement fraud analytics in imbalanced data environments.

The interpretability and imbalance sensitivity scores in Table II were derived from a synthesis of relevant literature (e.g., [29] and expert judgment based on the application of these metrics in fraud detection and procurement datasets. While not absolute, these heuristic scores facilitate a structured comparison of metric suitability in coalition fraud modelling. From the comparative analysis, it is evident that Confidence offers superior interpretability and is significantly less affected by class imbalance, an advantage that is particularly relevant in fraud detection settings where positive instances (i.e., fraud) are rare and scattered. Furthermore, its pattern-based nature enables it to surface relational anomalies, making it particularly well-suited for coalition detection.

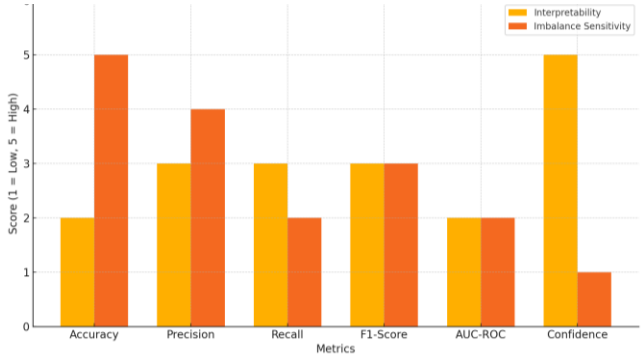


Fig. 1. Six evaluation metrics based on interpretability and imbalance sensitivity.

TABLE II. COMPARATIVE EVALUATION OF METRICS

No.	Metric	Interpretability (1–5)	Imbalance Sensitivity (1–5)	Justification
1.	Accuracy	2	5	Inflated in imbalanced datasets; does not reflect minority class detection [29]
2.	Precision	3	4	Useful to reduce false positives; appropriate in fraud context [30]
3.	Recall	3	2	Important when capturing all fraud instances; can result in high false positives [29]
4.	F1-Score	3	3	Balances precision and recall; effective for imbalanced data [31]
5.	AUC-ROC	2	2	Provides global model evaluation; lacks actionable insight for case-level interpretation [32]
6.	Confidence	5	1	Rule-based metric with high interpretability in pattern-based fraud; useful in coalition detection [22], [33]

C. Practical Insight: Rule-Based Detection

The primary advantage of Confidence lies in its ability to generate rule-level insights. For example, if a rule indicates that Supplier A and Supplier B co-bid in 90% of tenders where one is present, and both frequently win under similar conditions, this may trigger an alert. Such insight is difficult to derive from aggregate metrics like Accuracy or F1-score, which do not consider inter-entity relationships. Moreover, Confidence can be operationalized in conjunction with other rule-based measures such as support (frequency of occurrence) and lift (relative strength), enabling a multi-dimensional assessment of relational patterns. This is critical in the detection of cartel behavior, where [37] highlight the importance of using advanced analytical methods to distinguish genuine collusion from legitimate competition, a challenge made even more pressing in detecting cartel behavior where fraudulent patterns are often hidden within dense and repeated co-bidding networks.

D. Case-Based Justification

A simplified procurement case analysis drawn from Malaysian e-procurement data further underscores the utility of Confidence. These examples were derived from actual e-procurement datasets from Malaysian government agencies, with the dataset being on 2015, illustrating real-world bidding scenarios. In particular, multiple instances were identified where pairs of suppliers consistently co-bid with high Confidence scores (exceeding 80% conditional probability), strongly indicating potential collusive behavior. These relational patterns were not effectively flagged using traditional classification metrics due to their emphasis on class-labeled outcomes rather than inter-entity relationships.

The analysis revealed multiple instances where supplier pairs participated in tenders with statistically significant co-occurrence patterns. Traditional metrics failed to flag these cases due to their reliance on class-labeled outcomes, whereas Confidence-based rules identified relational anomalies worthy of further investigation. Table III illustrates the type, sensitivity to imbalanced data, interpretability, and use case in fraud detection for each metric.

TABLE III. TYPE, SENSITIVITY TO IMBALANCED DATA, INTERPRETABILITY, AND USE CASE IN FRAUD DETECTION FOR EACH METRICS

Metric	Type	Sensitivity to Imbalanced Data	Interpretability	Use Case in Fraud Detection	Author (s)
Accuracy	Overall performance	Low	High, but can be misleading	Not suitable for imbalanced datasets; may mask rare fraud cases	[15], [17]
Precision	Class-specific (Positive class)	High	High	Important to reduce false positives (e.g., avoid false alarms)	[17], [34]
Recall	Class-specific (Positive class)	High	High	Important to reduce false negatives (e.g., capture all fraud)	[17], [34]
F1-Score	Harmonic mean of precision & recall	Medium to High	Moderate	Balances precision and recall; useful in skewed data	[17], [34]
AUC-ROC	Threshold-independent	High	Moderate	Evaluates model across thresholds; compare classifier ability	[20]; [35]
Confidence	Rule-based (Association Rules)	High	High (in rule mining)	Measures rule strength; helps find fraud patterns	[23], [36]

E. Insight: Why Confidence is Useful for Coalition Detection

Confidence provides rule-level insight, which is particularly valuable for detecting procurement fraud coalitions. For example, a high confidence score for the co-occurrence of Supplier A and Supplier B (e.g., 90% confidence) suggests a strong relational pattern, which is more informative than an overall accuracy score when seeking hidden relationships. This characteristic supports early-warning systems and evidence-based procurement audits, aiding in identifying collusive behaviors that might not be apparent with aggregate metrics.

V. SYNTHESIS AND DISCUSSION

The comparative analysis presented in this review underscores the evolving challenges and opportunities in evaluating predictive models for procurement fraud detection, particularly to coalition or collusive behavior. While traditional evaluation metrics such as Accuracy, Precision, Recall, and AUC-ROC remain foundational in classification-based modeling, they are increasingly insufficient when applied to domains characterized by data imbalance, relational complexity, and pattern-level anomalies.

A. Limitations of Traditional Evaluation Metrics

A significant limitation of conventional metrics is their focus on instance-level classification performance. In procurement fraud detection, especially for coalition behaviors, the fraudulent signal often resides not in individual records but in the relationships among entities across time. For instance, Accuracy can be misleading in scenarios with extreme class imbalance, such as datasets where fraudulent tenders constitute less than 5% of all transactions by overstating model effectiveness due to the prevalence of true negatives [15], [29]. Similarly, although Precision and Recall offer greater sensitivity to the minority class, they remain inadequate in identifying the relational dynamics underlying collusion. F1-score, while useful in balancing both metrics, does not address interpretability or contextual relevance. AUC-ROC further abstracts the model's discriminatory power into a single score, which, although statistically sound, lacks the granularity needed for forensic or regulatory applications in procurement fraud.

B. Strengths of Confidence in Fraud Coalition Detection

The Confidence metric, grounded in association rule mining, provides a complementary and, in many cases, superior evaluative perspective for coalition fraud modeling. Unlike global metrics, Confidence quantifies the conditional likelihood of a co-occurrence, e.g., Supplier B submitting a bid when Supplier A does offering a highly interpretable and actionable view of potential collusion. Its unsupervised nature also makes it robust in contexts where labeled fraud data is scarce or incomplete. This is especially pertinent in the public sector, where investigations are often reactive and labels may only exist post-audit. The Confidence metric allows for early-warning signals through the detection of anomalous patterns, which can be subjected to further validation using domain expertise or additional metadata (e.g., company addresses, ownership structures, bid amounts). Importantly, Confidence should not be used in isolation. Its key limitation is the potential inflation of its value in low-support scenarios, where rare co-occurrences may appear statistically significant due to limited data. This issue can be mitigated by jointly analyzing support (absolute rule frequency) and lift (proportional deviation from independence), both of which enhance the reliability of Confidence-based findings [23].

C. Practical and Policy Implications

From a practical standpoint, integrating Confidence-based analytics into existing procurement monitoring systems could substantially enhance auditors' capabilities to identify relational anomalies indicative of collusive behavior. Policymakers and regulatory bodies, such as the Malaysia Competition Commission (MyCC), would benefit from adopting these interpretable analytics tools to improve transparency and proactively flag suspicious tenders. Implementing Confidence-based metrics can also inform risk-based auditing strategies and continuous supplier monitoring, ultimately promoting accountability and integrity in public procurement processes.

D. Implications for Practice and Policy

From a policy standpoint, the integration of Confidence into procurement analytics tools can inform risk-based auditing, automatic flagging of suspicious tenders, and ongoing supplier monitoring. As governments increase investment in e-procurement systems, embedding such interpretable analytics directly into workflows can improve the efficacy of oversight bodies like the Malaysia Competition Commission (MyCC) or anti-corruption agencies. For practitioners, the findings suggest a shift from solely performance-driven evaluation (i.e., how well does the model classify?) to insight-driven evaluation (i.e., what can the model explain about fraud structures?). This is aligned with broader trends in explainable AI (XAI) and human-centric data science.

E. Advancement Beyond Existing Studies

The comparative framework presented in this study extends prior research on evaluation metrics for fraud detection in three main ways. First, it introduces a cross-domain synthesis that integrates traditional classification metrics with rule-based interpretability measures such as Confidence an area rarely analyzed jointly in previous works. This dual perspective bridges the gap between performance evaluation and relationship discovery, enabling more meaningful assessments of models used in detecting collusion.

Second, by applying Confidence within the context of coalition fraud, the study advances the discussion from general fraud classification toward relational analytics, emphasizing inter-supplier dependencies that are typically overlooked by conventional measures like AUC-ROC or F1-score.

Third, the heuristic comparative matrix developed in this paper provides a transparent and structured tool for evaluating metrics along interpretability and class imbalance dimensions, which can be adopted by researchers and regulators for systematic assessment of fraud detection models.

Collectively, these contributions represent a methodological advancement over prior works that treated evaluation purely as a statistical exercise. The present study positions metric selection as both a technical and interpretive decision, thereby aligning predictive modeling more closely with real-world auditing and policy needs.

VI. Conclusion and Future Work

This review has critically examined the comparative utility of traditional classification-based evaluation metrics and the Confidence metric within the context of predictive modeling for procurement fraud, with a particular focus on coalition detection. The analysis demonstrates that while standard metrics such as Accuracy, Precision, Recall, F1-score, and AUC-ROC provide valuable insights for general model validation, they are insufficient when the modeling objective involves detecting rare, relational, and pattern-based fraud behaviors such as supplier collusion.

The Confidence metric, rooted in association rule mining, has been shown to offer superior interpretability, resilience to class imbalance, and applicability in unsupervised or semi-supervised settings. Its capacity to uncover rule-level relationships makes it particularly effective for modeling fraud coalitions, especially in scenarios where labeled data are scarce or incomplete. Despite its limitations, such as sensitivity to low-support rules, Confidence can be effectively calibrated using complementary metrics like support and lift. Given the unique strengths of both metric types, this paper advocates for the

adoption of hybrid evaluation frameworks that integrate classification metrics with rule-based measures. Such frameworks not only enhance model robustness but also contribute to explainability, operational relevance, and stakeholder trust, key considerations in public sector analytics and regulatory oversight.

Future work in procurement fraud analytics should focus on developing more intelligent and interpretable evaluation strategies to enhance both detection accuracy and real-world applicability. One key direction is the creation of composite scoring systems that integrate metrics such as confidence, support, lift, and conventional classification indicators (e.g., precision, recall, F1-score), enabling a more balanced trade-off between predictive performance and interpretability. Additionally, graph-based learning techniques, particularly graph neural networks (GNNs), offer promise in modeling coalition behavior by capturing structured relationships among suppliers; these models should be evaluated using hybrid metrics that reflect both statistical accuracy and semantic relevance.

Embedding confidence-based alerts into real-time procurement systems can further support dynamic risk monitoring and early intervention, aligning predictive insights with operational workflows. Furthermore, empirical validation using real procurement data, including blind evaluations by domain experts, is essential to assess the practical effectiveness and trustworthiness of confidence-driven rule discovery methods. Finally, conducting cross-national comparisons of procurement fraud detection frameworks will help generalize findings and identify best practices suited to varying legal and regulatory environments. As public procurement systems continue to evolve toward digitalization, the capacity to detect and explain collusive patterns will be central to ensuring transparency and preserving market integrity. This review contributes to that objective by outlining a roadmap for advancing interpretable, actionable, and context-sensitive evaluation of predictive models in fraud detection.

This study identified potential future initiatives to enhance the detection of procurement fraud, particularly in cases of collaboration. The thoughts are split into two groups: short-term and long-term.

A. Short-Term Priorities

First, it is a good idea to test the Confidence measure with more real data from purchases. A lot of articles solely utilize synthetic or example data. We can find out if Confidence can truly assist in uncovering collusion between suppliers if we utilize genuine tender data from a government or commercial system.

Second, researchers might be able to mix diverse measures. For example, the F1-score or Confidence with precision. This can help the model be both right and easy to grasp. In real life, people like auditors need to explain how the outcome came about, not just give a figure.

B. Long-Term Goals

Not all countries have the same system and data. Researchers ought to examine the functioning of Confidence across various countries. One strategy may work well in Malaysia, but it may not work the same way in other countries.

Finally, create a hybrid system that uses both rule-based and categorization metrics. This can make the model stronger and easier for others to comprehend and utilize. Auditors or officers can understand why the pattern is there and not merely follow the score.

Each metric was scored heuristically on a five-point Likert-type scale (1 = Low, 5 = High) for each dimension, drawing from synthesis of the literature and informed expert judgment. While such scoring does not provide definitive empirical quantification, it offers a structured and transparent means of articulating relative strengths and weaknesses among metrics. A common approach in methodological reviews is where direct comparisons are limited or context specific.

REFERENCES

[1] M. Kamal and A. Tohom, "Likelihood Rating of Fraud Risk in Government Procurement," The International Journal of Business Review (The Jobs Review), 2019, doi: 10.17509/tjr.v2i1.17903.

[2] D. M. Putri, D. Syariati, and A. Ahmad, "Fraud Prevention in the Perspective of Probity Audit (The Case Study of University X)," El Muhasaba Jurnal Akuntansi (E-Journal), 2021, doi: 10.18860/em.v12i2.12392.

[3] N. W. Rustiarini, T. Sutrisno, N. Nurkholis, and W. Andayani, "Why People Commit Public Procurement Fraud? The Fraud Diamond View," Journal of Public Procurement, 2019, doi: 10.1108/jopp-02-2019-0012.

[4] M. Gunasegaran, R. Basiruddin, and A. M. Rizal, "Detecting and Preventing Fraud in E-Procurement of Public Sector: A Review, Synthesis and Opportunities for Future Research," International Journal of Academic Research in Business and Social Sciences, 2023, doi: 10.6007/ijarbss/v13-i1/15970.

[5] N. Anggraeni, L. H. Husnan, and E. Pituringsih, "School Procurement Information System (SIPLah) as Moderating of Determinant of Fraud in Goods/Services Procurement," European Journal of Theoretical and Applied Sciences, 2023, doi: 10.59324/ejtas.2023.1(6).62.

[6] S. Sutoyo, M. F. Nugroho, and W. Windyastuti, "E-Procurement Implementation, Internal Control, and Organizational Commitment to Fraud Prevention Procurement of Goods and Servicesin Device Organizations Area (DOA) City Yogyakarta," Journal of Economics Finance and Management Studies, 2023, doi: 10.47191/jefms/v6-i5-54.

[7] N. W. Rustiarini, S. Sutrisno, N. Nurkholis, and W. Andayani, "Fraud Triangle in Public Procurement: Evidence From Indonesia," J Financ Crime, 2019, doi: 10.1108/jfc-11-2018-0121.

[8] A. P. Wicaksono, D. Urumsah, and F. Asmui, "The Implementation of E-Procurement System: Indonesia Evidence," SHS Web of Conferences, 2017, doi: 10.1051/shsconf/20173410004.

[9] Y. B. Adi and A. Rohman, "Determinants of Fraud Prevention of Procurement of Goods and Services in Government Agency," Jak (Jurnal Akuntansi) Kajian Ilmiah Akuntansi, 2023, doi: 10.30656/jak.v10i2.5498.

[10] E. Amwoha, "Constituents That Affect the Implementation of Sustainable Public Procurement in Kenyan Public Universities a Case of Technical University of Kenya," strategicjournals.com, 2015, doi: 10.61426/sjbcm.v2i1.78.

[11] N. Kumar and K. Ganguly, "Non-Financial E-Procurement Performance Measures," International Journal of Productivity and Performance Management, 2020, doi: 10.1108/ijppm-07-2019-0353.

[12] World Bank, "Common Red Flags of Fraud and Corruption in Procurement." [Online]. Available: https://documents1.worldbank.org/curated/en/223241573576857116/pdf/Warning-Signs-of-Fraud-and-Corruption-in-Procurement.pdf

[13] OECD, "ALGORITHMS AND COLLUSION - Competition policy in the digital age," 2017.

[14] Malaysia Competition Commission (MyCC), "Bid Rigging," 2022. Accessed: Jun. 13, 2023. [Online]. Available: https://www.mycc.gov.my/sites/default/files/pdf/newsroom/Bid%20Rigging%20BI_outline_0_0.pdf

[15] N. V Chawla, "Data mining for imbalanced datasets: An overview," in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, Eds., Springer, 2005, pp. 853–867. doi: 10.1007/0-387-25465-X_40.

[16] J. D. Kelleher, B. Mac Namee, and A. D'arcy, Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT press, 2020.

[17] V. Lopez, A. Fernandez, S. Garcia, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," Inf Sci (N Y), vol. 250, pp. 113–141, 2013, doi: 10.1016/j.ins.2013.07.007.

[18] 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.

[19] S. T. Boppiniti, "Machine learning for predictive analytics: Enhancing data-driven decision-making across industries," International Journal of Sustainable Development in Computing Science, vol. 1, no. 3, p. 13, 2019.

[20] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit Lett, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.

[21] M. Bowles, Machine Learning with Spark and Python: Essential Techniques for Predictive Analytics. John Wiley & Sons, 2019.

[22] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94), Santiago, Chile: Morgan Kaufmann, 1994, pp. 487–499. [Online]. Available: https://www.vldb.org/conf/1994/P487.PDF

[23] P.-N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining, 2nd ed. Pearson, 2019.

[24] A. P. Siddaway, A. M. Wood, and L. V Hedges, "How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses," Annu Rev Psychol, vol. 70, no. 1, pp. 747–770, 2019.

[25] M. Motevalli, "Comparative analysis of systematic, scoping, umbrella, and narrative reviews in clinical research: critical considerations and future directions," Int J Clin Pract, vol. 2025, no. 1, p. 9929300, 2025.

[26] M. Pautasso, "The structure and conduct of a narrative literature review," A guide to the scientific career: Virtues, communication, research and academic writing, pp. 299–310, 2019.

[27] J. Sukhera, "Narrative reviews: flexible, rigorous, and practical," J Grad Med Educ, vol. 14, no. 4, pp. 414–417, 2022.

[28] E. Toomey et al., "Focusing on fidelity: narrative review and recommendations for improving intervention fidelity within trials of health behaviour change interventions," Health Psychol Behav Med, vol. 8, no. 1, pp. 132–151, 2020.

[29] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Trans Knowl Data Eng, vol. 21, no. 9, pp. 1263–1284, 2009, doi: 10.1109/TKDE.2008.239.

[30] R. Saia and S. Carta, "Evaluating the benefits of using proactive transformed-domain-based techniques in fraud detection tasks," Future Generation Computer Systems, vol. 93, pp. 18–32, 2019.

[31] M. T. Islam, M. A. Rahman, M. T. R. Mazumder, and S. H. Shourov, "Comparative Analysis of Neural Network Architectures for Medical Image Classification: Evaluating Performance Across Diverse Models," American Journal of Advanced Technology and Engineering Solutions, vol. 4, no. 01, pp. 1–42, 2024.

[32] I. Michael, "Evaluation Metrics and Benchmarking for Predictive Accuracy," Jun. 2025.

[33] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explainable artificial intelligence in cybersecurity: A survey," Ieee Access, vol. 10, pp. 93575–93600, 2022.

[34] S. Dutta, P. Bhattacharya, and L. Dey, "An ensemble approach for early detection of banking fraud using soft computing techniques," Journal of Intelligent & Fuzzy Systems, vol. 31, no. 5, pp. 2689–2700, 2016, doi: 10.3233/IFS-162229.

[35] E. Lopez-Rojas and S. Axelsson, "Money laundering detection using synthetic data," in 2012 European Intelligence and Security Informatics Conference (EISIC), IEEE, 2012, pp. 252–259. doi: 10.1109/EISIC.2012.53.

[36] R. Agrawal, T. Imielinski, and A. Swami, "Mining association rules between sets of items in large databases," in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Association for Computing Machinery, 1993, pp. 207–216. doi: 10.1145/170035.170072.

[37] C. Carbone, F. Calderoni, and M. Jofre, "Bid-rigging in public procurement: cartel strategies and bidding patterns," Crime Law Soc Change, vol. 82, no. 2, pp. 249–281, 2024.

Correcting Blue-Shift in Single-Image Dehazing via Haze-Compensated Von Kries Adaptation

Asniyani Nur Haidar Abdullah[1], Mohd Shafry Mohd Rahim[2], Sim Hiew Moi[3],
Azah Kamilah Draman[4], Ahmad Hoirul Basori[5], Novanto Yudistira[6]

Pervasive Computing & Educational Technology, Center for Advanced Computing Technology (C-ACT), Department of Media Interactive-Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia [1]

Department of Emergent Computing-Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia[2]

Department of Software Engineering-Faculty of Computing, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia[3]

Department of Software Engineering, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka, Malaysia[4]

Department of Information Technology-Faculty of Computing and Information Technology in Rabigh King Abdulaziz University, Rabigh, Saudi Arabia[5]

Departemen Teknik Informatika-Fakultas Ilmu Komputer, Universitas Brawijaya, Jl. Veteran, Ketawanggede, Lowokwaru, Kota Malang, Jawa Timur, Indonesia – 65145[6]

Abstract*—Haze severely degrades image quality by reducing contrast, obscuring details, and introducing a blue-shift color cast caused by atmospheric scattering. Traditional dehazing methods, including prior-based approaches (e.g., DCP, CAP, LPMinVP) and preprocessing techniques (e.g., ICAP WB, Dynamic Gamma), improve visibility but fail to correct haze-induced color imbalance, resulting in unstable RGB distributions and unnatural tone reproduction. This study proposes the Haze-Compensated Color Von Kries (HCCVK) method, a lightweight and training-free preprocessing strategy that performs color compensation before transmission estimation in single-image dehazing. HCCVK integrates a novel red-channel compensation mechanism with Von Kries chromatic adaptation to mitigate wavelength-dependent haze suppression and stabilize chromatic consistency under varying illumination. Unlike learning-based color correction approaches, HCCVK does not require training data, is computationally efficient, and maintains algorithmic interpretability, making it suitable for practical deployment. The method was evaluated on six benchmark datasets: CHIC, Dense-Haze, I-Haze, O-Haze, SOT, and NH-Haze, covering indoor, outdoor, dense, and non-homogeneous haze scenarios. Experimental results based on the RGB color balance metric (σRGB) show that HCCVK reduces color deviation by approximately 75–92% on CHIC, 80–90% on Dense-Haze, and 82–90% on NH-Haze compared to the widely used DCP, and also outperforms CAP, ICAP WB, Dynamic Gamma, and LPMinVP by producing more compact and stable RGB distributions. These findings demonstrate that HCCVK effectively corrects blue-shift imbalance, preserves luminance consistency, and enhances the color stability of dehazing pipelines.

Keywords*—Image dehazing; blue-shift correction; color compensation; Von Kries adaptation; preprocessing

I. Introduction

Haze is an atmospheric phenomenon that degrades image quality due to the scattering and absorption of light by particles such as water droplets, dust, and smoke suspended in the air [1]. This degradation reduces contrast, obscures fine details, and introduces a characteristic blue-shift color cast caused by wavelength-dependent scattering, where shorter wavelengths (blue) are scattered more strongly than longer wavelengths (red) [2][3]. Such distortions negatively affect human perception and degrade the performance of computer vision applications, including autonomous driving, UAV imaging, security surveillance, and remote sensing [4][5].

Similarly, studies in other computer vision domains have shown that poor quality input data leads to degraded feature extraction and reduced system performance. For example, Waheed et al. [6] demonstrated that melanoma classification accuracy declines when dermoscopic images are not properly preprocessed before CNN analysis, while Awan et al. [7] emphasized that preprocessing significantly reduces execution time and improves classification stability in MRI-based diagnosis. Furthermore, Jabal et al. [8] showed that structural errors in raw CAD drawings lead to error propagation during feature extraction, reinforcing the importance of early corrective processing. Therefore, haze removal is essential for both visual enhancement and robust scene understanding under real-world outdoor conditions [9-10].

Single-image dehazing techniques are commonly categorized into prior-based, preprocessing-based, and learning-based methods. Prior-based methods rely on physical or statistical assumptions to estimate scene transmission and atmospheric light. Examples include the Dark Channel Prior (DCP) [11] and the Color Attenuation Prior (CAP) [12], including their enhanced variants ICAP [13], Extremum Value prior-based Local Patch Wise Minimal Values (LPMinVP) [14] and Dynamic Gamma [15]. These methods can enhance structural visibility but often introduce color distortions, halo artifacts, and inconsistent illumination recovery in dense haze regions due to inaccurate atmospheric light estimation [13]. Preprocessing-based techniques such as white balancing, histogram equalization, and gamma correction improve global illumination but do not explicitly correct haze-induced color

bias, leading to unstable RGB distributions [14][16]. Recently, deep learning methods such as CCD-Net [17] and attention-guided restoration models [18] have achieved significant improvements, yet they depend on large-scale training datasets, high computational cost, and lack interpretability, limiting their deployment in resource-constrained or real-time applications.

A major limitation across existing dehazing methods is that color imbalance caused by atmospheric scattering is not explicitly corrected before transmission estimation. Due to wavelength-dependent scattering, the red channel experiences greater attenuation than the blue channel, resulting in a dominant blue-shift color cast in hazy images [3][19]. When dehazing is applied directly to such chromatically unbalanced inputs, error propagation occurs during transmission estimation, leading to unnatural tone reproduction and unstable RGB restoration [13][14][17]. This type of error propagation from unbalanced inputs has also been highlighted in medical image analysis and CAD extraction research, indicating that effective preprocessing is necessary before advanced processing [8].

Therefore, this study addresses the lack of a lightweight, training-free preprocessing mechanism that compensates red-channel attenuation and stabilizes RGB color channels before dehazing. To solve this problem, we propose the Haze-Compensated Color Von Kries (HCCVK) method, a preprocessing strategy that integrates red-channel compensation with Von Kries chromatic adaptation to correct haze-induced spectral imbalance. Unlike deep learning–based enhancement approaches that require GPU training resources, HCCVK requires no training data [7], offers full interpretability, and maintains low computational cost, making it suitable for real-time deployment.

II. Related Works

Single-image dehazing techniques have been widely explored in computer vision and can be categorized into three main groups: prior-based methods, preprocessing-based methods, and learning-based methods. Prior-based dehazing methods estimate scene transmission based on statistical assumptions about haze-free images. The DCP [11] assumes that at least one color channel in a local image patch has low intensity, enabling effective haze removal. However, it frequently generates halo artifacts and color distortion, especially in bright regions and sky areas where the prior is invalid. The CAP [112] utilizes the relationship between depth and brightness to estimate scene depth but often suffers from over-saturation and instability under varying illumination. Improved variants such as ICAP [13] and LPMinVP [14] enhance transmission estimation using adaptive priors but still struggle with chromatic inconsistency and residual color bias, particularly in dense haze conditions. Overall, prior-based methods enhance visibility but inherently lack a mechanism to correct haze-induced color imbalance.

Preprocessing-based approaches aim to enhance visibility before dehazing by normalizing illumination or contrast. White balance correction is commonly used to suppress color cast but does not correct wavelength-dependent red-channel suppression in haze [16]. Similarly, gamma correction and histogram equalization improve brightness and contrast but often amplify noise and cause over-enhancement [14]. Methods such as Dynamic Gamma (DD-Gamma) [15] improve global tone distribution but still fail to address blue-shift caused by atmospheric scattering, resulting in unstable RGB distributions. These approaches improve perceptual quality but do not provide physically meaningful color correction.

Learning-based dehazing methods have recently gained popularity by leveraging convolutional neural networks to learn end-to-end haze removal. Models such as AOD-Net and DCPDN incorporate physical priors into CNN architectures, while CCD-Net [17] introduces a dual-branch fusion network in RGB and Lab color space to reduce perceptual distortions. Other studies employ transformer-based attention or adversarial learning to improve scene restoration [18]. However, these methods generally require large training datasets, lack interpretability, and perform poorly in real-world atmospheric variations due to training bias and overfitting. Additionally, most learning-based models do not explicitly solve the color imbalance problem and often rely on post-processing color correction.

The importance of early-stage preprocessing to prevent feature distortion has also been emphasized across other computer vision fields. For example, Waheed et al. demonstrated that CNN-based melanoma classification can only achieve high accuracy when color distortion and illumination are corrected during preprocessing [6]. Similarly, Awan et al. showed that preprocessing improves both diagnostic stability and computational efficiency in MRI image analysis [7]. Additionally, Jabal et al. reported that failure to correct noisy or inconsistent CAD input data leads to error propagation in feature extraction systems [8]. These findings support the need for preprocessing strategies that stabilize input data before complex restoration or estimation steps.

Based on this review, it is evident that most dehazing research prioritizes visibility enhancement but neglects chromatic fidelity. Neither prior-based methods nor preprocessing and learning strategies provide a reliable solution to haze-induced blue-shift and red-channel attenuation, which leads to unnatural tone reproduction and unstable color restoration. Unlike these methods, the proposed HCCVK method focuses specifically on stabilizing RGB balance before dehazing, offering a lightweight and interpretable solution to mitigate color distortion caused by atmospheric scattering.

III. Methodology

This section outlines the experimental design, including the proposed HCCVK method, benchmark datasets, and evaluation metrics. The methodology is structured to ensure reproducibility and clarity for future implementations. It also demonstrates the effectiveness of the proposed approach under a variety of indoor and outdoor hazy conditions. Color balance is applied as a preprocessing step to mitigate blue-channel bias and preserve luminance consistency under hazy conditions.

Preprocessing plays a crucial role in preventing feature degradation during downstream processing, especially when the input data suffers from structural or spectral distortion. Similar to dehazing challenges, Waheed et al. reported that

classification accuracy in melanoma skin cancer detection drops significantly if illumination bias and color distortion are not corrected prior to CNN classification [6]. Likewise, Awan et al. demonstrated that MRI-based diagnostic performance improves when input instability is reduced through intensity normalization and contrast enhancement, especially in resource-constrained environments [7]. In another domain, Jabal et al. emphasized that uncorrected noise in CAD drawings propagates errors in feature extraction and interpretation, highlighting the importance of robust preprocessing before complex analysis [8]. Motivated by these findings, the HCCVK method introduces a lightweight preprocessing strategy to stabilize color distribution before dehazing, preventing chromatic error propagation during transmission estimation.

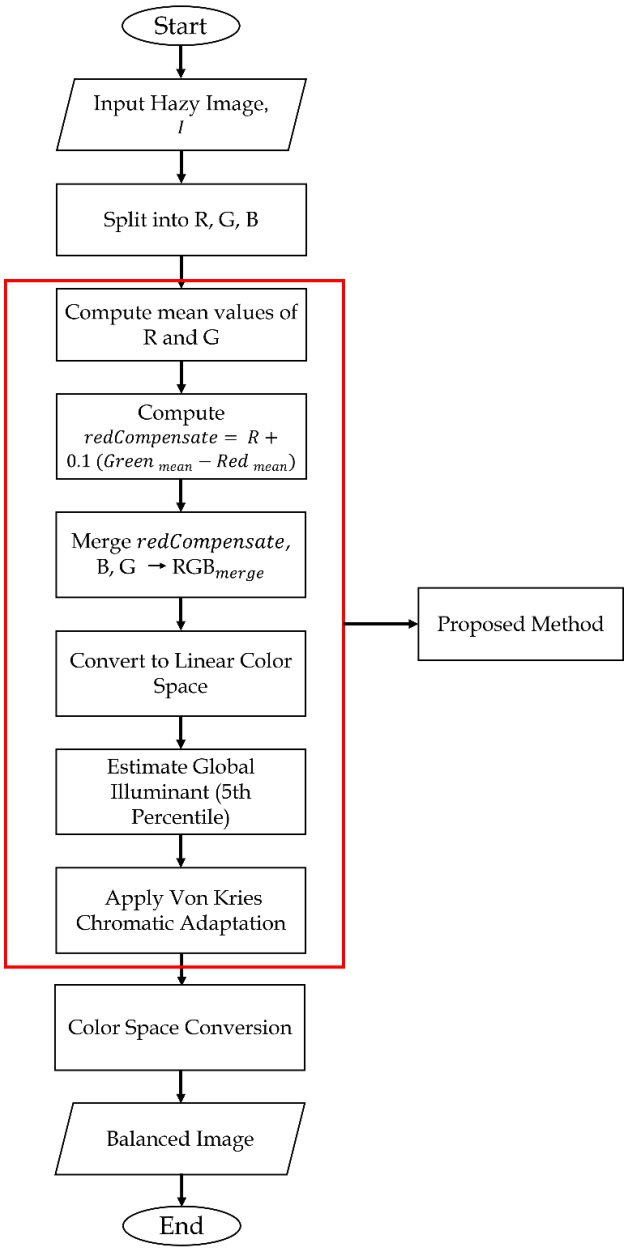


Fig. 1. Flowchart of the proposed Haze-Compensated Color Von Kries method for haze-compensated color balance.

A. Proposed Haze-Compensate Color Von Kries Method Acronyms

The HCCVK method is designed as a preprocessing stage prior to dehazing. It balances the RGB distribution by compensating the red channel and correcting the global illuminant through the Von Kries chromatic adaptation transformation. Unlike deep learning–based preprocessing techniques, which often require dedicated training resources and GPU acceleration [7], HCCVK is computationally efficient and entirely training-free. The overall workflow of the method is illustrated in Fig. 1.

As shown in the flowchart, the method proceeds through a sequence of operations from channel separation to illuminant correction. For clarity and reproducibility, the detailed steps are presented in Algorithm 1.

Algorithm 1 Haze-Compensated Color Von Kries (HCCVK) Method

Input: Hazy Image, I

Output: Haze-compensated balanced image, $I_{balance}$

Step:

1) Preprocessing:
   a) Split the input image into RGB channels: R, G, B.
   b) Compute mean intensities: $Red_{mean}, Green_{mean}$
   c) Apply red-channel compensation: $redCompensate = R + 0.1\,(Green_{mean} - Red_{mean})$

2) Reconstruction
   a) Merge $redCompensate$, G, and B to obtain RGB_{merge}
   b) Convert RGB_{merge} into linear color space: RGB_{linear}

3) Illuminant Estimation:
   a) Define a 5% percentile threshold for robust estimation
   b) Estimate global illuminant using the gray-world assumption at the 5th percentile

4) Chromatic Adaptation:
   a) Apply the Von Kries transform using the estimated illuminant
   b) Generate corrected image $I_{balance}$

5) Finalization:
   a) Convert $I_{balance}$ back to RGB color space
   b) Output the haze-compensated balanced image

As shown in Table I, the weight of 0.1 achieved the highest average compensation and the lowest standard deviation among all tested parameters. This finding was further validated through entropy analysis, where entropy is defined as:

$$e = -\sum_{i=0}^{255}\sum_{j=0}^{255}\sum_{k=1}^{255} p_{i,j,k}.log_2(p_{ijk}) \quad (1)$$

Here, i,j,k denote the three dimensions of the probability distribution. The analysis was conducted over 390 images selected from the total dataset of 1,239, covering both controlled (indoor) and non-controlled (outdoor) illumination

conditions. Therefore, a fixed weighting factor of 0.1 was adopted in the proposed method to ensure stability and robustness across different haze conditions.

TABLE I. WEIGHTED RED AND BLUE CHANNEL COMPENSATION COMPARISON WITH ENTROPY INFORMATION

Weight	Entropy Information, e			
	Average		Standard Deviation	
	Channel			
	Red	Blue	Red	Blue
0.1	**6.8503**	6.8452	**0.4738**	0.4762
0.3	6.7879	6.7847	0.5229	0.5263
0.5	6.7873	6.7690	0.5248	0.5384
0.7	6.8288	6.7976	0.4852	0.5071
0.9	6.7739	6.7456	0.5355	0.5644

B. Dataset

To ensure a comprehensive evaluation and demonstrate the robustness of the proposed method, a total of 1,239 hazy images with corresponding ground-truth (GT) images were utilized. The dataset comprises 588 indoor images and 651 outdoor images, enabling performance assessment under different environmental conditions. These variations ensure that the evaluation covers both controlled scenarios and more challenging natural conditions.

Indoor datasets were captured under controlled illumination, which guarantees stable lighting and minimizes external variability. These include I-Haze with 30 images [20], SOT from the RESIDE dataset with 500 images [21], CHIC with 36 images [22], and Dense-Haze with 22 images [23]. Such conditions make them suitable for analyzing the performance of the method in consistent environments.

The outdoor datasets were collected under natural illumination, where changes in weather, solar position, and atmospheric scattering introduce greater complexity for dehazing. These include O-Haze with 45 images [20], SOT with 500 images [21], CHIC with 18 images [22], Dense-Haze with 33 images [23], and NH-Haze with 55 images [19]. The inclusion of these datasets ensures that the method is evaluated across both controlled and highly variable conditions.

C. Evaluation Matrix

The effectiveness of the proposed HCCVK method was evaluated using the RGB Color Balance metric (σRGB), which quantifies the degree of consistency among the three color channels. This metric is particularly suitable for haze removal tasks, as hazy conditions often suppress the red channel, leading to visible color imbalance. A lower σRGB value indicates better channel consistency and improved color fidelity, while higher values reflect imbalance and potential color casts. The metric is defined as:

$$\sigma RGB = \sqrt{\frac{(R-\mu)^2+(G-\mu)^2+(B-\mu)^2}{3}}, \mu = \frac{R+G+B}{3} \quad (2)$$

where R, G, and B denote the average intensities of the red, green, and blue channels, and μ represents their mean intensity. Unlike structural indices such as PSNR and SSIM, which emphasize pixel-level or textural similarity, σRGB directly measures the preservation of color balance, aligning more closely with the objectives of the proposed method.

D. Comparative Analysis

For benchmarking, the proposed method (PM OUT) was compared against several well-established dehazing algorithms, including DCP, CAP, ICAP WB, LPMinVP, DD GAMMA, PM CC and PM CM. These methods were chosen because they represent both prior-based and preprocessing-driven strategies, ensuring a fair and comprehensive evaluation. Since each approach directly affects brightness preservation and color balance, their inclusion enables a systematic comparison, highlighting not only overall dehazing performance but also the influence of distinct preprocessing modules on final image quality.

IV. RESULTS

This section evaluates the performance of the proposed HCCVK method using the σRGB, where lower values indicate better chromatic stability and reduced blue-shift distortion. HCCVK is compared against five baseline methods: DCP, CAP, ICAP WB, LPMinVP, and Dynamic Gamma. Results are reported across six benchmark datasets covering indoor, outdoor, dense, and non-homogeneous haze conditions.

A. CHIC Dataset (Indoor and Outdoor)

The results in Fig. 2 show that on the CHIC Indoor dataset, HCCVK consistently achieves the lowest σRGB values among all methods, demonstrating stable color restoration across varying illumination. While LPMinVP and ICAP WB partially improve color consistency, their σRGB ranges remain significantly wider than HCCVK, indicating lingering blue-shift imbalance. DCP exhibits the highest σRGB deviation due to its inaccurate atmospheric light estimation in bright regions.

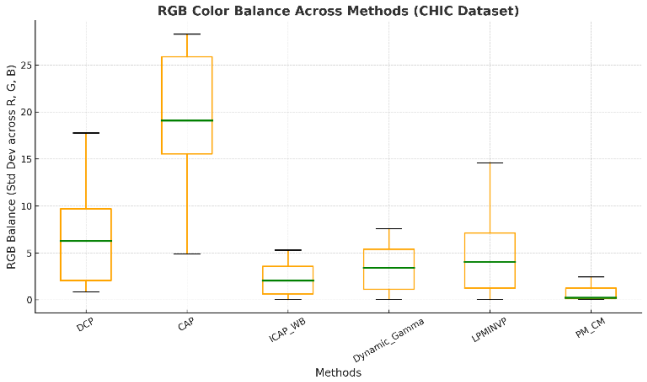


Fig. 2. RGB color balance comparison across dehazing methods on the CHIC indoor dataset.

For the CHIC Outdoor dataset (Fig. 3), similar behavior is observed. HCCVK maintains a compact σRGB distribution, confirming its effectiveness in both indoor and outdoor haze environments. Competing methods such as CAP and Dynamic Gamma struggle to stabilize the red and green channels, resulting in color shifts in scenes with strong sky illumination. In comparison, HCCVK corrects red-channel suppression, producing visually natural color tone balance.

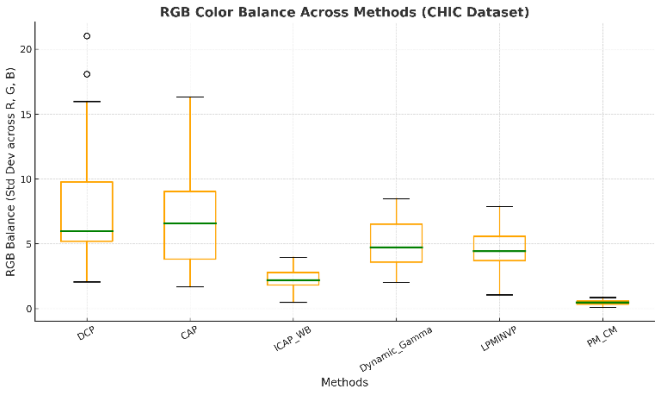


Fig. 3. RGB color balance comparison across dehazing methods on the CHIC outdoor dataset.

B. Dense Dataset (Indoor and Outdoor)

On the Dense-Haze dataset (Fig. 4, 5), haze density increases image color distortion. Despite this challenge, HCCVK maintains the most stable RGB balance with minimal variance. Competing methods such as LPMinVP outperform DCP and CAP in high-density haze but still suffer chromatic instability, shown by higher σRGB interquartile ranges. This indicates that intensity equalization alone is insufficient without spectral compensation.

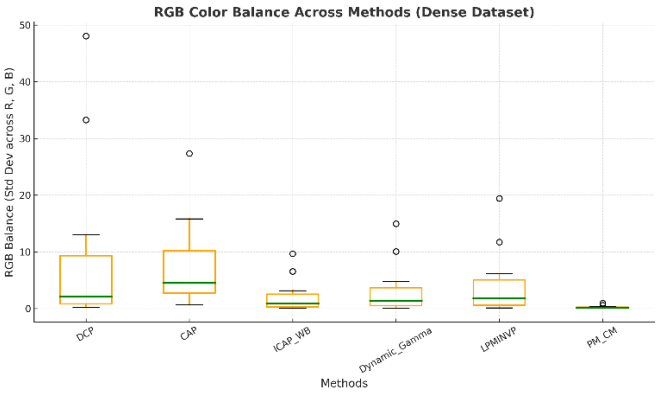


Fig. 4. RGB color balance comparison across dehazing methods on the dense indoor dataset.

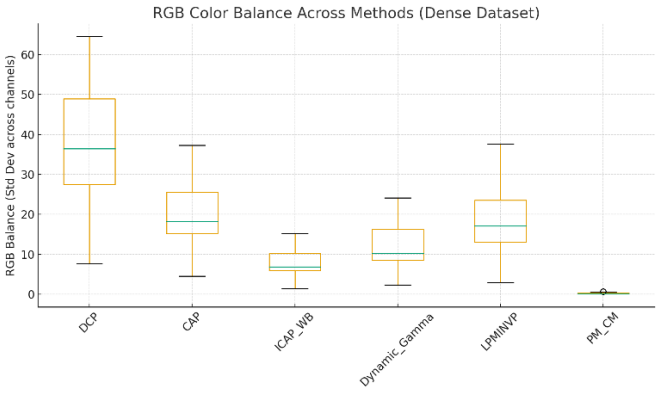


Fig. 5. RGB color balance comparison across dehazing methods on the dense outdoor dataset.

C. I-Haze and O-Haze Dataset

The I-Haze and O-Haze datasets (Fig. 6–7) evaluate haze removal in real historical scenes. HCCVK consistently produces lower σRGB compared to learning-free baselines, proving strong generalization to real-world atmospheric haze. ICAP WB improves luminance consistency but fails to correct haze-induced red-channel attenuation, leading to residual blue tint in several scenes. HCCVK effectively addresses this limitation.

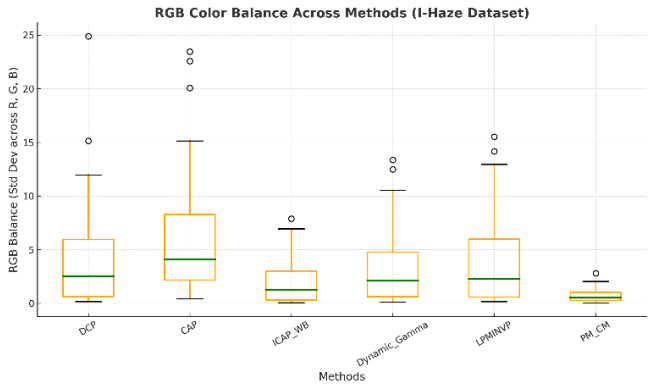


Fig. 6. RGB color balance comparison across dehazing methods on the I-Haze dataset.

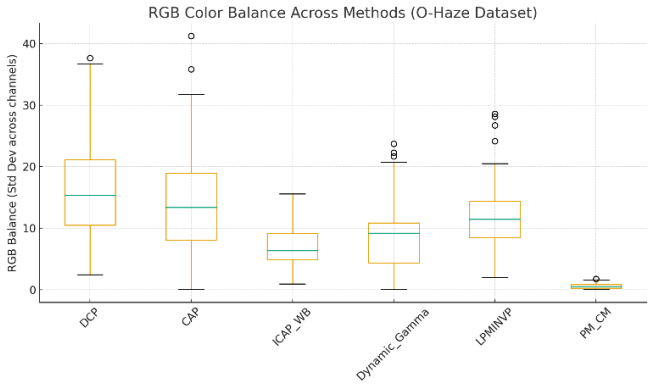


Fig. 7. RGB color balance comparison across dehazing methods on the O-Haze dataset.

D. SOT Dataset (Indoor and Outdoor)

On the SOT dataset (Fig. 8–9), which contains challenging texture transitions and depth variations, HCCVK provides the tightest σRGB dispersion, reflecting robust channel stability. CAP and Dynamic Gamma show inconsistent behavior, performing well in some scenes but failing in low-light regions due to over-enhancement artifacts.

E. NH Dataset

Finally, the NH-Haze dataset (Fig. 10), representing non-homogeneous haze, poses the greatest difficulty. HCCVK again achieves the lowest color imbalance and strongest robustness against haze variation. Prior-based methods degrade sharply in this dataset, demonstrating their limited adaptability to haze thickness variation. In contrast, HCCVK maintains consistent σRGB even under severe haze scattering.

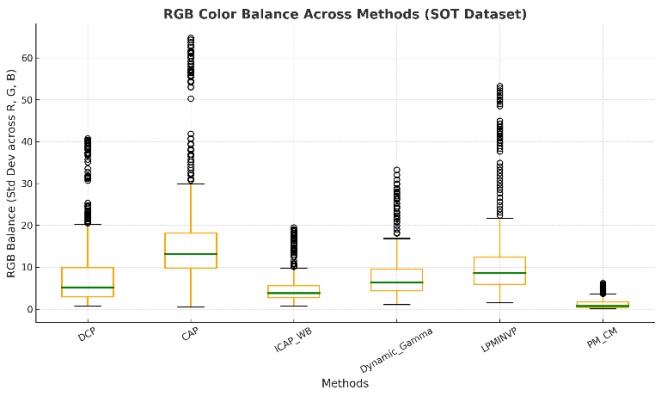


Fig. 8. RGB color balance comparison across dehazing methods on the SOT indoor dataset.

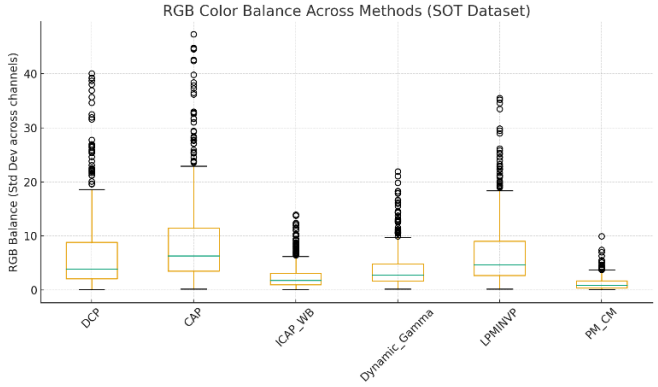


Fig. 9. RGB color balance comparison across dehazing methods on the SOT outdoor dataset.

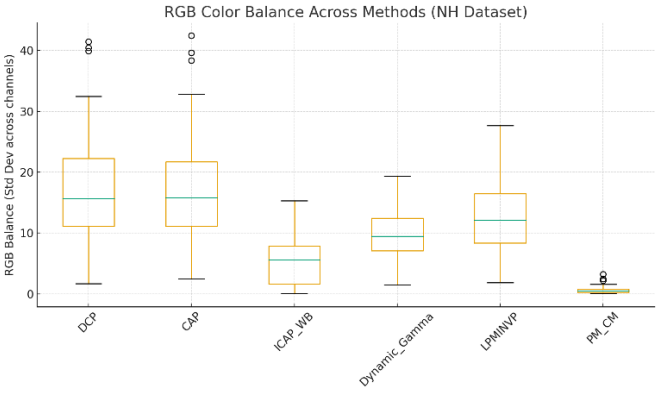


Fig. 10. RGB color balance comparison across dehazing methods on the NH dataset.

V. Discussion

The experimental results across six benchmark datasets consistently demonstrate that the proposed HCCVK method outperforms both prior-based and preprocessing-driven dehazing techniques. The superiority of HCCVK can be attributed to two core design principles: 1) explicit compensation of the red channel to counteract haze-induced blue-shift imbalance, and 2) global illuminant correction through Von Kries chromatic adaptation. These steps stabilize RGB distributions before dehazing, yielding more compact box plots and lower σRGB values compared to competing methods.

Prior-based methods such as DCP and CAP exhibit high variability and frequent outliers n σRGB, indicating unstable color restoration, especially under severe haze. These observations are consistent with findings reported in previous studies. Guo et al. [24] showed that classical scattering-based priors are sensitive to atmospheric light estimation errors, often causing excessive color deviation and halo artifacts. Similarly, Ancuti et al. [19] emphasized in the NTIRE dehazing challenge that handcrafted priors struggle to preserve color fidelity, especially in outdoor haze scenarios. Our results reinforce these limitations, as DCP and CAP show wide σRGB dispersion across most datasets, demonstrating their inability to correct haze-induced color imbalance.

Preprocessing-driven methods such as ICAP WB, LPMinVP, and Dynamic Gamma show moderate improvement in color consistency but remain inadequate for haze compensation. These methods apply intensity or gamma corrections without addressing wavelength-dependent scattering, resulting in incomplete chromatic recovery. Husain and Rahim [15] observed that gamma-based adjustments improve brightness but fail to eliminate residual blue-shift, while Huang et al. [25] found that simple color balancing lacks robustness in non-uniform haze conditions. This limitation is evident in our results, where ICAP WB and LPMinVP produce higher median σRGB and broader channel variations compared to HCCVK.

Compared to recent learning-based approaches, HCCVK maintains strong generalization without requiring model training. Chen et al. [17] introduced CCD-Net to address color distortion in deep networks, while Jing et al. [18] implemented color correction modules to improve stability under changing illumination. Although effective, these methods rely on large annotated datasets and intensive computation, limiting real-time applicability. In contrast, HCCVK delivers competitive color correction without training, offering interpretability and low computational cost.

Overall, the consistently narrow σRGB range obtained by HCCVK across CHIC, Dense-Haze, I-Haze, O-Haze, SOT, and NH-Haze confirms its robust chromatic recovery and blue-shift mitigation capability. Unlike existing methods that treat color correction as a post-processing step, HCCVK stabilizes chromatic information before dehazing, preventing color error propagation. This makes HCCVK a generalizable and deployment-ready preprocessing module for haze removal, especially in real-world applications such as intelligent transportation, visual navigation, and UAV imaging.

VI. Conclusion and Future Work

This study presented the HCCVK method, a lightweight preprocessing solution designed to correct haze-induced chromatic imbalance before dehazing. By integrating a red-channel compensation mechanism with Von Kries chromatic adaptation in linear RGB space, the method directly addresses the spectral attenuation that leads to blue-shift dominance in hazy images. Experimental validation on six heterogeneous benchmark datasets (CHIC, Dense-Haze, I-Haze, O-Haze, SOT, and NH-Haze) demonstrated that HCCVK consistently achieved the lowest σRGB values and the most compact channel distributions, indicating superior chromatic stability

compared to prior-based and preprocessing-driven baselines. These results confirm that addressing color imbalance at the preprocessing stage prevents chromatic error propagation during dehazing and provides a more reliable foundation for subsequent visibility restoration.

Beyond performance, the significance of HCCVK lies in its interpretability, generalizability, and low computational cost, making it suitable for deployment in real-time and resource-constrained computer vision systems. Unlike learning-based approaches, it does not require training or dataset-dependent tuning, yet maintains consistent robustness across varying haze densities and illumination conditions.

Future work will extend this research in three directions. First, HCCVK will be integrated with existing deep learning dehazing networks to explore hybrid learning-enhancement architectures. Second, additional evaluation metrics such as CIEDE2000, SSIM, and perceptual metrics like LPIPS will be incorporated to further quantify visual quality improvements. Finally, the method will be expanded to video dehazing and multimodal fusion scenarios to support safety-critical applications such as autonomous driving, UAV navigation, maritime vision, and intelligent surveillance.

ACKNOWLEDGMENT

This work was supported by Universiti Teknikal Malaysia Melaka (UTeM) under research grant PJP/2024/FTMK/PERINTIS/SA0043. The authors would also like to acknowledge the continuous support from the Center for Research and Innovation Management (CRIM), UTeM.

REFERENCES

[1] M. Kaur, D. Singh, V. Kumar, U. Rawat, & M. Amoon, "Dsscnet: deep custom spatial and spectral consistency layer-based dehazing network", IEEE Access, vol. 12, p. 44325-44334, 2024.

[2] K. Hu, Q. Zeng, J. Wang, H. Jianqing, & Y. Qi, "A method for defogging sea fog images by integrating dark channel prior with adaptive sky region segmentation", Journal of Marine Science and Engineering, vol. 12, no. 8, p. 1255, 2024.

[3] S. Yoon and J. Cho, "Deep multimodal detection in reduced visibility using thermal depth estimation for autonomous driving", Sensors, vol. 22, no. 14, p. 5084, 2022.

[4] T. Liu and Z. Baijun, "Dual-channel and two-stage dehazing network for promoting ship detection in visual perception system", Mathematical Problems in Engineering, vol. 2022, p. 1-15, 2022.

[5] X. Li, Z. Qiao, G. Wan, S. Zhu, Z. Zhongxin, X. Fan, et al., "Depth-guided bilateral grid feature fusion network for dehazing", Sensors, vol. 24, no. 11, p. 3589, 2024.

[6] S. R. Waheed, S. M. Saadi, M. S. M. Rahim, N. M. Suaib, F. H. Najjar et al., "Melanoma skin cancer classification based on CNN deep learning algorithms," Malaysian Journal of Fundamental and Applied Sciences, vol. 19, no. 3, pp. 299–305, 2023.

[7] M. J. Awan, M. S. M. Rahim, N. Salim, A. Ismail, and H. Shabbir, "Acceleration of knee MRI cancellous bone classification on Google Colaboratory using convolutional neural network," Int. J. Adv. Trends Comput. Sci., vol. 8, pp. 83–88, 2019.

[8] M. F. A. Jabal, M. S. M. Rahim, N. Z. S. Othman, and Z. Jupri, "A comparative study on extraction and recognition method of CAD data from CAD drawings," in Proc. 2009 Int. Conf. Information Management and Engineering, pp. 709–713, IEEE, 2009.

[9] X. He, T. Jia, & J. Li, "Learning degradation-aware visual prompt for maritime image restoration under adverse weather conditions", Frontiers in Marine Science, vol. 11, 2024.

[10] D. Lee, J. Im, & J. Won, "Virtual lidar sensor intensity data modeling for autonomous driving simulators", IEEE Access, vol. 11, p. 120694-120706, 2023.

[11] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, 2010.

[12] Q. Zhu, J. Mai, and L. Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3522–3533, 2015.

[13] D. Ngo, G. D. Lee, and B. Kang, "Improved color attenuation prior for single-image haze removal," *Appl. Sci.*, vol. 9, no. 19, art. no. 4011, 2019.

[14] J. Han, J. Liu, and Y. Li, "Extremum value prior-based dehazing method," Signal Processing: Image Communication, vol. 99, p. 116469, Feb. 2022.

[15] A. Husain and M. S. M. Rahim, "Dynamic gamma correction-based single image dehazing," *IEEE Access*, vol. 10, pp. 72410–72423, Jul. 2022.

[16] C. Kim, "Region adaptive single image dehazing," Entropy, vol. 23, no. 11, p. 1438, 2021.

[17] X. Chen, J. Li, and L. Wang, "CCD-Net: A color-correction dehazing network based on dual-branch fusion," *Applied Sciences*, vol. 15, no. 6, p. 3191, 2025.

[18] H. Jing, P. Liu, and S. Zhou, "An efficient dehazing method using pixel unshuffle and color correction," *Signal Processing: Image Communication*, vol. 125, p. 117203, 2025.

[19] C. Ancuti, C. Ancuti, and R. Timofte, "NTIRE 2024 dense and non-homogeneous dehazing challenge report," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, Jun. 2024, pp. 123–138.

[20] C.-O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-HAZE: A dehazing benchmark with real hazy and haze-free outdoor images," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 2018, pp. 754–762.

[21] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, et al., "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, Jan. 2019.

[22] J. El Khoury, J.-B. Thomas, and A. Mansouri, "A database with reference for image dehazing evaluation," *Journal of Imaging Science and Technology*, vol. 62, no. 1, pp. 10503–10515, Jan. 2018.

[23] C.-O. Ancuti, C. Ancuti, M. Sbert, and R. Timofte, "Dense-HAZE: A benchmark for image dehazing with dense-haze and haze-free images," in *Proc. IEEE Int. Conf. Image Processing (ICIP)*, Taipei, Taiwan, Sept. 2019, pp. 1014–1018.

[24] Y. Guo, J. Wang, and H. Zhang, "Single image dehazing method based on Rayleigh scattering and adaptive adjustment," *PLoS One*, vol. 20, no. 5, e0315176, 2025.

[25] W. Huang, H. Zhou, J. Du, and Y. Chen, "Single image dehazing based on color balancing and quad-decomposition," *Optics & Laser Technology*, vol. 163, p. 109441, 2023.

Enhancing Dermatological Diagnostics: An Enhanced Approach for Skin Cancer Classification Using pix2pix GAN

Adnan Afroz[1], Shaheena Noor[2], Shakil Ahmed Bashir[3], Umair Jilani[4]
Dept. of Computer Engineering, Sir Syed University of Engineering and Technology, Karachi, Pakistan[1]
Dept. of Computer Engineering, Sir Syed University of Engineering and Technology, Karachi, Pakistan[2]
Dept. of Computer Science, FAST National University, Karachi, Pakistan[3]
Dept. of Telecommunication Engineering, Sir Syed University of Engineering and Technology, Karachi, Pakistan[4]

***Abstract*—Skin cancer is among the predominant forms of the disease that includes malignant squamous cell carcinoma, basal cell carcinoma, and melanoma that is characterized by aberrant melanocyte cell development. Frequent screenings and examinations enhance the prognosis for people with skin cancer. Sadly, a lot of patients with skin cancer are not diagnosed until the condition has progressed past the point at which treatment is effective. Deep learning techniques in computer vision have made impressive strides, but issues like class imbalance and a lack of data still hinder the autonomous identification of skin conditions. A solution to address these problems is the implementation of GAN, which is capable of synthesizing realistic data. In this paper, a deep learning GAN model for image synthesis utilizing the pix2pixHD integrated with Convolutional Neural Network (CNN) classifier approach is used to perform skin cancer classification. To categorize three forms of skin cancer benign or malignant. The proposed pix2pixHD GAN is a novel method for utilizing pertinent skin lesion information for generation of high-quality synthesized dermoscopic image and conduct skin lesion classification performance with improved accuracy. Realistic images were created using a U-Net-based generator and PatchGAN discriminator with custom CNN architecture to classify three forms of cancer. With remarkable accuracy of 87.65% (MEL), 91% (BCC), and 89.85% (SCC) and other performance parameters indicate that GAN pix2pixHD Classifier model has promising results in classification. These findings demonstrate the Classifier's ability to produce and correctly identify high-quality skin lesion images, indicating its potential as a deep learning-based medical image analysis tool.**

***Keywords*—Deep learning; skin cancer; generative adversarial network; pix2pixHD; classification**

I. Introduction

The human body's most important and sensitive organ, the skin, acts as the body's main defense against the environment and is essential for shielding the body from dangerous UV rays and illnesses. But because it is the outermost layer, it can get a variety of ailments, including skin cancer, which is one of the most prevalent cancers in the world [1] Many people are suffering from different types of skin cancer. In the year 2021 there were approximately 6.64 million new cases of skin cancers worldwide, encompassing both melanoma and non-melanoma types [2]. The stats provided that in year 2022, an estimated 330,000 new cases of melanoma were diagnosed globally, with nearly 60,000 deaths attributed to the disease [3].

As per study, in the US in 2019, about 100,000 new cases was found and 7,000 people will die. Lucky for us, melanoma treatment works much better when it was found early; it has a 92% 5-year survival rate, but only 26% when it was found later. So, finding and diagnosing melanoma as soon as possible is very important[4].

Current improvements in computer vision using Artificial Intelligence (AI) have improved diagnosis and detection approaches for fatal malignant skin conditions. Skin cancer diagnosis has always relied on manual screening and physical inspection. Dermatologists' procedures for visual inspection and evaluating of disease images are challenging, unreliable, and mistake prone[4]. The cancerous pixels must be recognized correctly to do investigation, interpretation, and comprehension of skin lesion images. This is challenging for the reasons stated below.

- Hair, veins, lubricants, bubbles, and various other clutter may be present in lesions on the skin images, interfering with the segmentation process.
- A lack of contrast between the adjacent skin and the lesion area also makes precise segmentation of the lesion difficult.
- Skin lesions typically vary in size, form, and color, limiting the efficiency of procedures in obtaining a greater accuracy degree level.

Skin lesions typically vary in size, form, and color, limiting the efficiency of procedures in obtaining a greater accuracy degree level. Due to this, demonologist find manual diagnosis challenging and necessitates the implementation of computerized diagnostic for analysis of skin lesion in making fast decisions. It has helped specialists make more precise and rapid recommendations concerning but its most threaten type Melanoma, have a very high death rate, but they are easy to treat if they are found early [4].

Over the years, different automated methods have been used to find and diagnose skin cancer. This is how medical imaging analysis is generally done: by using various dermoscopic image processing techniques. In general, the complete procedure of

finding melanoma cancer can be broken down into four essential steps:

- Preliminary image processing.
- Segmentation of images.
- Extraction of features.
- Classification of cancerous tumors.

Some of the other techniques used in preliminary processing include adjusting contrast and sharpness, morphological operations, binarization, data augmentation and colour grey-scaling. Noise and other irregularities are taken out of images at this stage. Resizing images makes sure that they are all the same size, which also makes computations easier. Segmentation of skin images run right after pre-processing, of images, separates the diseased area from the healthy meat to get the region of interest.

It divides healthy tissue prior to collecting lesions' features for suitable diagnosis. Skin lesion analysis for the identification of melanoma uses thresholds, clustering, edge and region-based, and ABCDE segmentation Algorithm. Such methods struggle to successfully partition the Melanoma region because of the complicated visual appearance of the skin lesions [5].

The majority of current research on skin cancer classification uses either traditional CNN architectures or GAN-based data augmentation for binary classification, often with poor image quality and insufficient management of data imbalance. This research fills this gap by combining a CNN classifier with pix2pix GAN to improve lesion representation and diagnostic performance for BCC, SCC, and melanoma.

II. Methods

A. Skin Cancer Diagnosis Through Deep Learning

The fine-grained variety of skin lesion appearances makes the task of classifying skin lesions difficult. As high-performance computing has advanced and massive datasets have emerged, deep learning techniques have demonstrated significant promise in a variety of image processing applications. Additionally, the most deployed deep learning convolutional neural network (CNN) performs end-to-end training using labels and image pixel input. CNN is better than the hand-crafted method in automatically extracting salient characteristics from the input data. CNN used for the first time in 1995 and utilized for analysis of medical imaging [6]. To train a deep CNN model with strong generalization capabilities, sufficient labelled data is needed. However, there aren't many publicly accessible data points for classifying skin lesions. The model transfer technique through transfer education are the main categories of classification [7].

The 129,450 clinical images were used to fine-tune the GoogleNet Inception v3 CNN architecture, which had been pretrained on the ImageNet dataset. With 2,032 disease types serving as its leaf nodes, and tree-structured taxonomy-based dataset is arranged. A partitioning algorithm is used to define the training classes based on these leaf nodes. Two binary classification challenges were used to evaluate the CNN model's performance with that of 21 dermatologists. The findings demonstrate that in certain skin cancer classification tasks, artificial intelligence has surpassed human experts. Numerous follow-up researchers have been drawn to this seminal work [8].

Researcher looked into using artificial intelligence and dermatologists to classify skin lesions [9]. The authors generated two sets of classification results using a ResNet50 model [10]. The XGBoost method was then used to merge the findings. Artificial intelligence and human decision-making together demonstrated superiority over separate systems [11].

Research has been conducted and demonstrated that augmented data is beneficial for accuracy by accurate classification and preventing overfitting by training a 4-layered CNN model with to distinguish benign skin lesions from melanoma using both normal and supplemented datasets [12].

By training two dual deep CNNs and allowing them to support one another, a unique skin lesion categorization model was created [12]. The ISIC 2016 Challenge dataset yielded state-of-the-art results for the suggested model [13].

Data synthesis, a more advanced technique for data augmentation, is intriguing and well-liked. Variational autoencoders (VAEs) [14] and generative adversarial networks (GANs) are the two most used methods for data synthesis [15]. In addition to classifying the input data into distinct classes, the WGAN can produce 64 x 64 images. The suggested approach used just 140 labelled images from the ISIC 2016 Challenge dataset, yielding an average precision score of 0.424 [16].

Research was conducted and presented a deeply differentiated GAN (DDGAN) with the goal of improving the resolution of produced image [17]. In order to demonstrate that DDGAN [18] and LAPGAN [19] that both can learn the distributed datasets and synthesize samples that looks like realistic, the authors created realistic-looking 256 x 256 skin lesion image and compared them.

GAN for image synthesis using pix2pixHD. They synthesize maps and semantic maps rather than creating images from random noise. The suggested pix2pixHD GAN is a unique approach to using the relevant skin lesion information for synthesizing high-quality image , and it is somewhat conducive to skin lesion classification performance, even if it needs annotated data to produce images, but this approach only characterized two types of skin cancer benign or melanoma [20].

B. Dataset

We require certain masks that demonstrate the presence or absence of clinically significant skin lesion patterns to train and evaluate pix2pix.

These masks may be found among the 2,601 images from the 2018 ISIC training dataset[21]. The segmentation masks for the same lesions, which are utilized to create instance and semantic maps, were acquired via the 2018 ISIC Challenge[21]. To train 2,056 images and test 545 images, we divided the data for three different classes. The details of divided data available in Table I.

A resilient adversarial learning objective function, a patch GAN discriminator architecture, and a generator with U-net are used in the pix2pix network that is a conditional image-to-image

translation GAN. The network was able to operate with high-resolution samples as shown in Fig. 1.

Differentiating between Basal Cell Carcinoma, Melanoma and Squamous Cell Carcinoma is in skin categorization process. Conventional CNN-based classifiers use used that classify in between three dermoscopic images to directly learn features.

Pix2pix model is used to generate synthetic images and trained to improve dermoscopic images as shown in Fig. 2(a) and 2(b). The CNN classifier learn features from these generated images and classify the three skin cancer types. CNN classifier has total to 115,232 trainable parameters, max pool and fully connected layer at output.

TABLE I. DATA FOR TRAINING AND TESTING

Image Class	Training	Testing	Total
Basal Cell Carcinoma(BCC)	679	188	867
Melanoma(MEL)	700	167	867
Squamous Cell Carcinoma(SCC)	677	190	867
Total	2056	545	2601

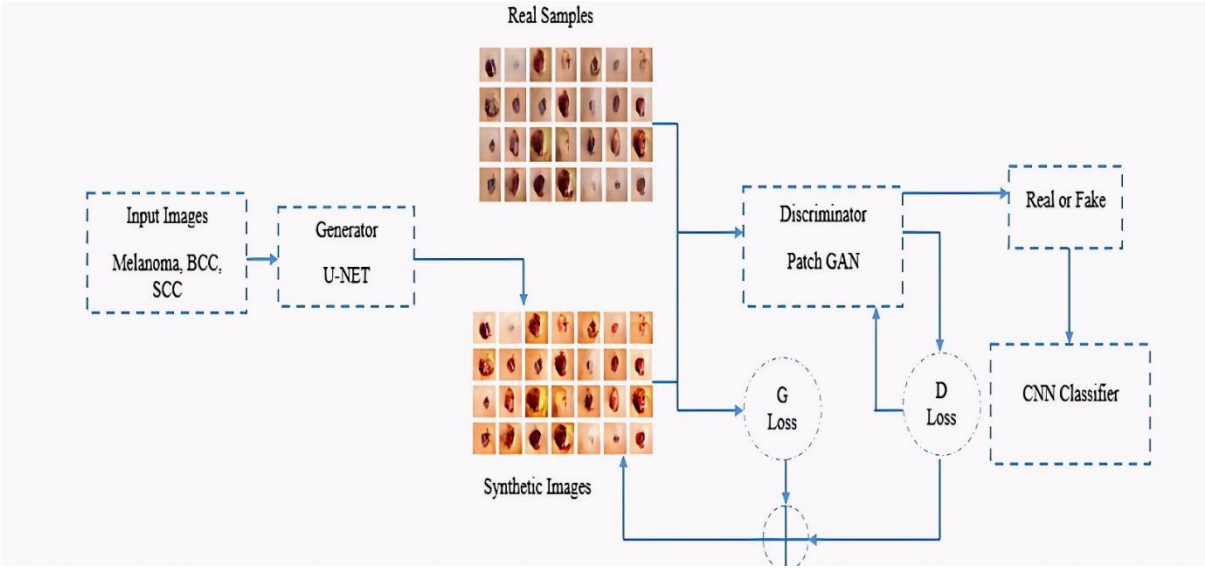
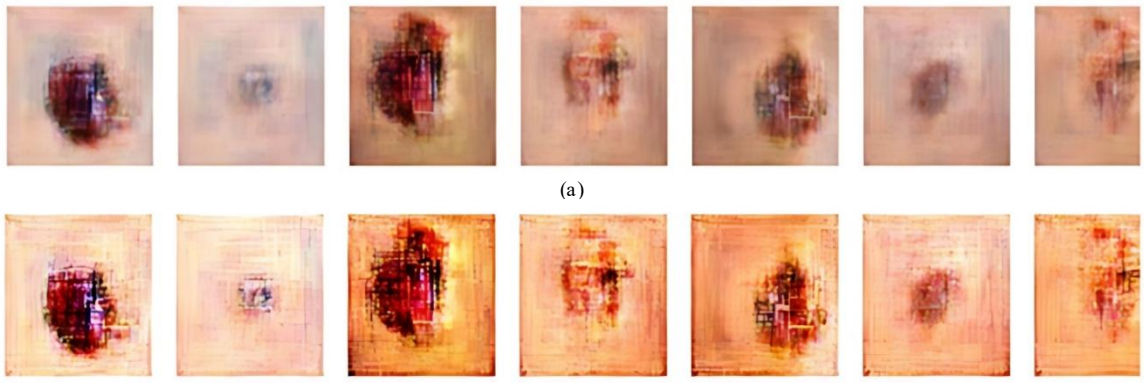


Fig. 1. Pix2pix GAN: Image generation and classification.



(a)

(b)

Fig. 2. (a) Original image input from ISIC 2018 dataset, (b) High resolution synthetic images generated from pix2pix GAN.

A U-Net design, which is useful for maintaining spatial information essential for medical image, is commonly used by the pix2pix generator. The discriminator is a PatchGAN, which helps the model concentrate on texture and fine features by classifying tiny patches of the Image instead of the entire image. Adversarial loss and L1 loss the pixel-by-pixel difference between the produced Image and the ground truth are used in tandem to train the pix2pix model. Segmentation masks or diagnostic labels can be used to direct the learning process for melanoma tasks. After training, the pix2pix output is utilized as a preprocessing step for an independent classification network or directly for diagnosis. In certain instances, the produced image alone can be sufficiently interpreted to help doctors make diagnoses.

The generator is encouraged to provide outputs that are identical to actual images by the adversarial loss as shown in Eq. (1):

$$LcGAN(G,D) = Ex,y[logD(x,y)] + Ex,z[log(1 - D(x,G(x,z)))] \quad (1)$$

An L1 loss as shown in Eq. (2) is used to guarantee that the produced images near the ground truth that improve pixel level accuracy and decrease blurriness.

$$\lambda \mathcal{L}_{L1}(G) = \mathrm{E}x,y[\| y - G(x) \| 1] \quad (2)$$

Realism is guaranteed by the GAN loss. The input's fidelity is guaranteed by the L1 loss. When combined, they yield results that are correct semantically and believable visually as shown in Eq. (3):

$$G^* = \arg min_G max_D \; L_C \; GAN(G,D) + \lambda \mathcal{L}_{L1}(G) \quad (3)$$

The model architecture consists of generator with U-Net attention, input of RGB image of 256 x 256, activation function of ReLU in encoder, LeakyReLU in decoder, and Tanh at final output as illustrated in Fig. 3.

The U-Net architectural design is for the purpose of medical image segmentation. It has a symmetric structure that constitutes an encoder and decoder: the encoder will collect contextual information through down-sampling, while the decoder restores fine-grained details by up-sampling. A distinguishing feature that gives U-Net its advantage is the presence of skip connections between symmetric layers of the encoder to the decoder. These connections help in retaining some spatial information that could be lost during down-sampling and make U-Net particularly good at tasks that need high localization accuracy, such as boundary delineation in medical images [22].

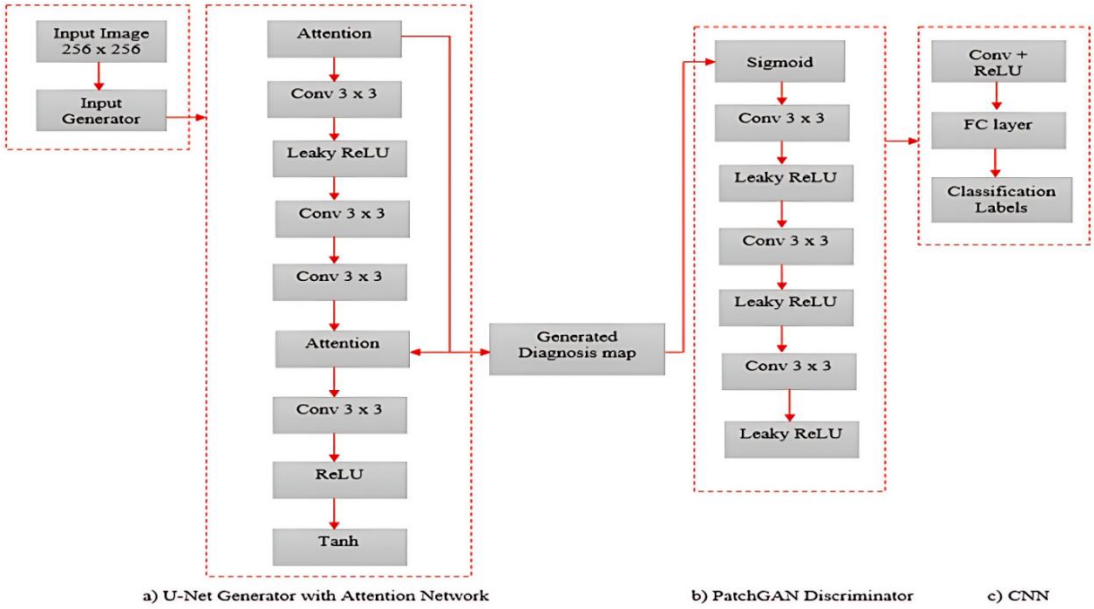


Fig. 3. Pix2pix GAN architecture with U-Net generator and CNN as classifier.

The U-Net architecture acts as a generator in a pix2pix GAN setup developed for medical-image-segmentation tasks. First, dermoscopic images (256×256 pixels) are fed into the encoder-decoder structure. The encoder successively down-samples the image, using convolutional layers with ReLU activation and batch normalization (except for the last layer), extracting deeper features while reducing spatial resolution. With each down-sampling step, the number of feature channels is increased so that the model can learn complex patterns. The decoder upsamples these features to reconstruct the image at the original resolution using transposed convolutions and LeakyReLU activations. A vital design consideration is the use of skip connections between corresponding encoder and decoder layers so that fine-grained features from the earlier layers can directly influence the output. Such preservation of spatial details minimizes information loss in the down-sampling process. Finally, the single-channel output (256×256×1) representing the diagnosis or segmentation map is activated using the Tanh function, constraining pixel values to the range of [-1,1]. This U-Net generator is pivotal to attaining realistic and spatially accurate outputs in the pix2pix GAN pipeline [22]. The inclusion of attention in this way improves both interpretability of the model and its segmentation accuracy, especially in applications dealing with skin cancer imaging. The model is therefore facilitated in providing more accurate and realistic segmentations of lesions in aiding clinicians in suspicious area identification. Visualizing these attention maps presents additional insight into what the model considers most relevant for diagnostic purposes.

Linear transformation performed in Attention network achieved through encoder feature map in Eq. (4):

$$x \in R^{(Fx \times H \times W)} \quad (4)$$

The decoder feature map is achieved through gating signal along with skip connections to produce feature map in Eq. (5).

$$g \in R^{(Fg \times H' \times W')} \quad (5)$$

Applying the linear projections on both equations through convolution shown in Eq. (6) and (7).

$$\theta_x = W_{(x)} x \quad (6)$$

$$\phi_g = W_{(g)} g \quad (7)$$

Two inputs are projected, added and passed through activation function ReLU that helps in learning the non-linear projections between encoder and decoder as shown in Eq. (8).

$$\psi=\text{ReLU}\,(\theta_x+\phi_g) \tag{8}$$

Scalar map is needed to produce attention network for knowing the important spatial regions by using sigmoid as activation function and convolution network as shown in Eq. (9).

$$\alpha=\sigma(W_\psi\ \psi+b_\psi) \tag{9}$$

The discriminator usually uses a PatchGAN architecture, The discriminator usually uses a PatchGAN architecture, which works as a Markovian discriminator by assessing the realism of tiny, overlapping patches 70×70 rather than the full image for balanced realism & accuracy. Because it enforces for high accuracy, this localized technique is especially well-suited for medical imaging jobs where fine-grained characteristics like lesion boundaries, pigmentation, and texture changes are crucial for diagnosis. In order to encourage the generator to generate anatomically correct and spatially coherent outputs, PatchGAN slides a neural classifier over the Image and classifies each patch as real or false. This results in more realistic lesion shapes and clearer borders, both of which are necessary for accurate skin cancer segmentation and detection. The Adam Optimizer is used with tuned parameters. We tried different sets of values for different parameters but there is problem of fast convergence. The values of β1 and β2 are set to stabilize the training process as available in Table II:

TABLE II. TUNED PARAMETER FOR PIX2PIX GAN

Parameter	Value
Input Image Size	256x256
Batch Size	16
Learning Rate (Generator)	0.0002
Learning Rate (Discriminator)	0.0001
Optimizer	Adam
β1	0.5
Epochs	200

III. PERFORMANCE EVALUATION

Classification purposes in the health domain wants great sensitivity as it signifies the measure of the system's value. Sensitivity is stated as Eq. (10):

$$Sensitivity/Recall = \frac{(TP)}{(TP+FN)} \tag{10}$$

Precision represents the ratio of accurately labels that are classified truly positive, which is stated as Eq. (11):

$$Precision = \frac{(TN)}{(FP+TN)} \tag{11}$$

Accuracy calculates the sum of appropriately classified skin lesions divided by the total number of skin lesions, which is stated as Equation 12 and F1-score in Eq. (13):

$$Accuracy = \frac{(TP+TN)}{(TP+FN)+(FP+TN)} \tag{12}$$

$$F1\ Score = \frac{(2*TP)}{(2*TP+FN+FP)} \tag{13}$$

IV. RESULTS AND DISCUSSION

We train a skin classification network with both synthetic and actual training sets to assess the entire set of synthetic images. Then, we compare the area images. For this comparison, we employ three distinct synthetic images created with our version of pix2pix that use both instance and semantic maps are called instances.

Only semantic label maps are used to produce semantic samples, whereas our conditional PGAN is used to generate PGAN samples. The best synthetic samples are those that are produced using instance maps. The findings for synthetic images confirm that they include characteristics that indicate whether a lesion is MEL, BCC and SCC. Furthermore, the findings imply that the created images have characteristics that go beyond what is found in the actual images. The predicted output is shown Fig. 4 that has three classified outputs.

The performance evaluation for our implemented CNN classifier is available in Table III. These values confirm that the model maintains strong precision-recall performance, especially for BCC, with slightly more variability in MEL and SCC. The implemented model reveals that the performance evaluators are nearly equal to accuracy for predicted classes.

TABLE III. PERFORMANCE EVALUATORS FOR PIX2PIX GAN AND CNN AS CLASSIFIER

Class	Accuracy	Precision	Recall	F1-Score
BCC	91%	93%	91%	92%
MEL	87.65%	89.18%	88.06%	88.1%
SCC	89.85%	87.23%	90%	88%

In order to validate the result, we have implemented the dataset on different deep learning models and the results are available in Table IV. For different models the recall, F1 and precision values are close to accuracy for VGG-16 MobileNet-V2 and InceptionV3 model that depicts the accuracy of generated synthetic images.

TABLE IV. PERFORMANCE EVALUATORS FOR DIFFERENT DEEP LEARNING MODELS

	VGG-16[23]			
Class	**Accuracy**	**Recall**	**F1-Score**	**Precision**
BCC	84.83%	83.27%	83.42%	83.62%
MEL	82.18%	79.83%	78.30%	80.63%
SCC	83.78%	81.00%	79.00%	80.31%
	MobileNet-V2[24]			
BCC	79.58%	79.30%	78.20%	80.97%
MEL	84.63%	82.02%	77.23%	82.89%
SCC	82.58%	80.28%	80.00%	81.00%
	InceptionV3[25]			
BCC	82.13%	80.45%	82.64%	80.27%
MEL	83.02%	81.75%	82.33%	82.42%
SCC	85.89%	85.00%	83.97%	83.14%

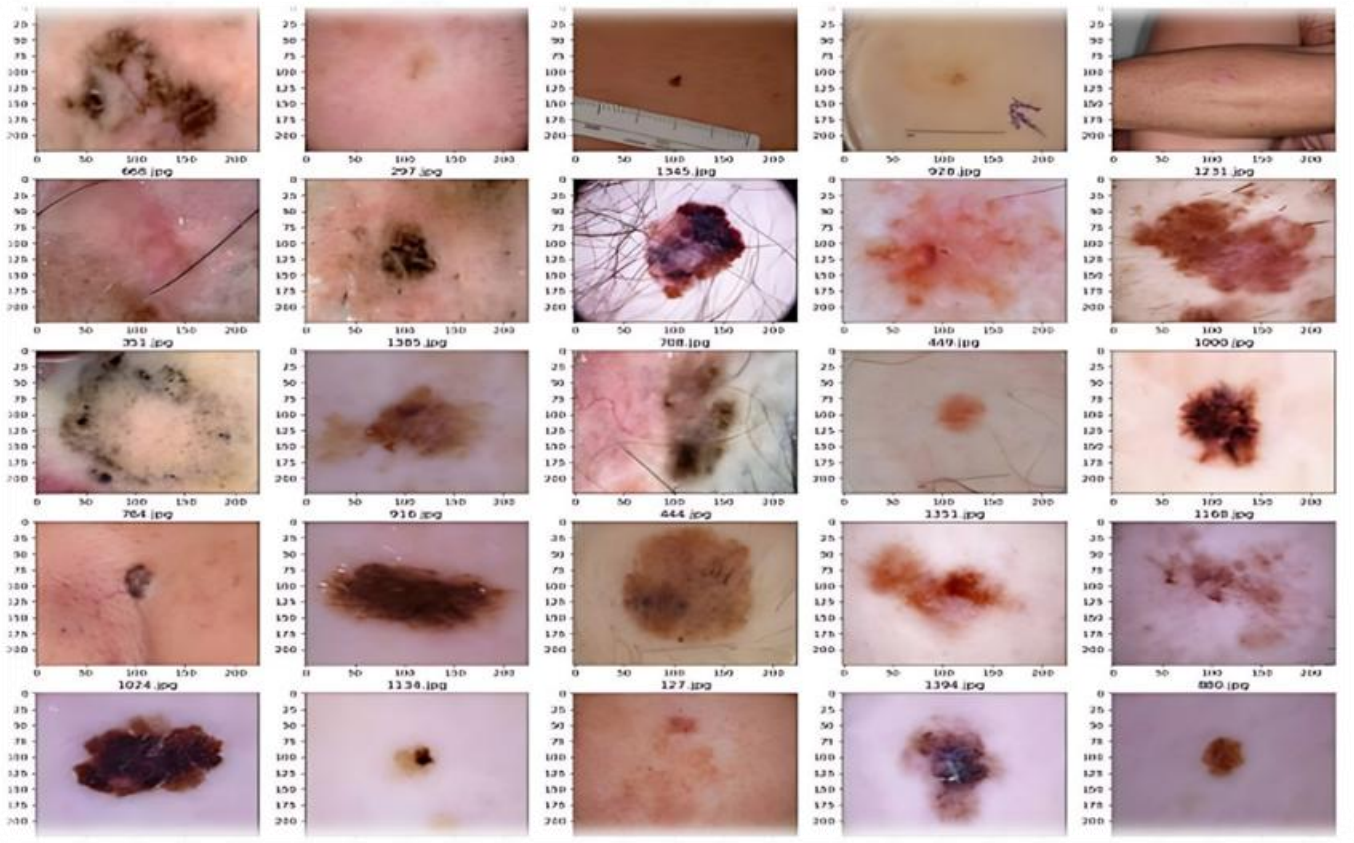


Fig. 4. Three classified outputs after CNN classification.

The model was trained on CPU 7th Generation core i-7 Intel processor, with 16 gigabytes (GB) of NVDIA GeForce GTX 1080 with 8 GB of RAM make up the system's Graphical Processing Unit (GPU). The confusion matrix for our implemented CNN classifier for three types of skin cancer for classification is shown in Fig. 5.

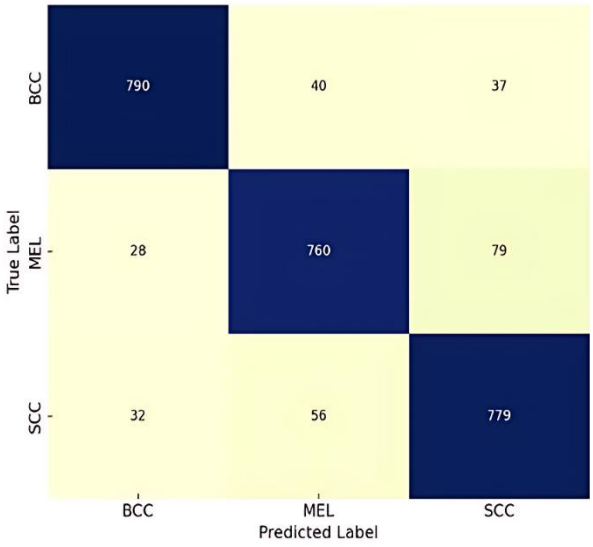


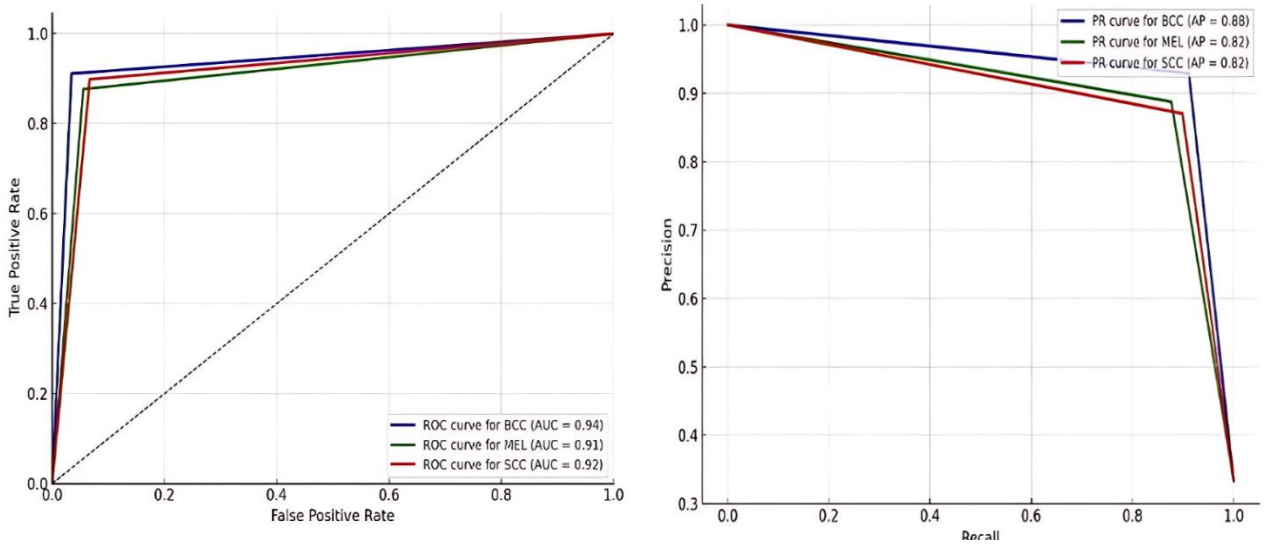
Fig. 5. Confusion matrix for classification of MEL, BCC and SCC through CNN classifier.

In order to validate the result, we have implemented the dataset on different deep learning models and the results are available in Table IV. ROC and Precision-Recall curves for MEL, BCC and SCC are shown in Fig. 6(a) and Fig. 6(b). For different models the recall, F1 and precision values are close to accuracy for VGG-16, MobileNet-V2 and InceptionV3 model that depicts the accuracy of generated synthetic images.

Results clearly shows that our proposed hybrid method, implementing pix2pix GANs with CNNs, predicts 91% accuracy on BCC, 87.65% on MEL and 89.85% on SCC using the ISIC-2018 dataset compared with typical CNN models like VGG-16, MobileNet-V2, and InceptionV3 by roughly 5 to 10% with reduced computational power and enhanced multi-class skin lesion classification. Table V also displays the comparison-based classification performance accuracy results of other authors' related work.

TABLE V. CLASSIFICATION PERFORMANCE ACCURACY FOR EXISTING RESEARCH

Author	Model	Classes	Accuracy
[8]	Inception v3 CNN	2	85%
[9]	Hybrid	2	88%
[12]	Synergic CNN	2	82-85%
[17]	MelanoGANs	2	<85
[20]	pix2pix GAN	2	85



(a) (b)

Fig. 6. (a) ROC curves for MEL, BCC and SCC, (b) Precision-recall curves for MEL, BCC and SCC.

V. Conclusion

Medical imaging is rapidly being driven by hybrid deep learning frameworks that combine generative and discriminative models. This research work used a pix2pix GAN with a U-Net generator and PatchGAN discriminator, together with a customized CNN classifier, to improve the classification of melanoma (MEL), basal cell carcinoma (BCC), and squamous cell carcinoma (SCC). The model efficiently controlled data imbalance and variability by synthesizing high-resolution dermoscopic images, reaching classification accuracies of 91% (BCC), 87.65% (MEL), and 89.85% (SCC), outperforming other models such as VGG-16, MobileNet-V2, and InceptionV3. These findings highlight the potential for GAN-augmented classification to increase diagnostic performance and robustness. A strong transition toward clinically deployable, high-precision diagnostic systems in dermatology is shown by this, which is consistent with current research trends that emphasize data augmentation and multimodal learning.

VI. Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of the paper.

References

[1] L. Wei, Zhang, Y., & Liu, J, "Enhanced skin cancer diagnosis using optimized CNN architecture," BMC Medical Imaging, vol. 24, no. 1, pp. 15-25, 2024.

[2] Z. Y. Zhou L, Han L, Xie Y, Wan M. Global, "Regional, and national trends in the burden of melanoma and non-melanoma skin cancer: insights from the global burden of disease study 1990-2021," Sci. Rep vol. 15, no. 1, 2025, doi: doi: 10.1038/s41598-025-90485-3.

[3] I. A. f. R. o. Cancer. "IARC marks Global Non-Melanoma Skin Cancer Awareness Day." https://www.iarc.who.int/cancer-type/skin-cancer (accessed.

[4] R. Gupta, & Sharma, "Skin cancer detection and classification using neural networks: A systematic review," Journal of Engineering and Computer Innovations, vol. 5, no. 2, pp. 45-60, 2024.

[5] S. S. G. Pellacani, "Comparison between morphological parameters in pigmented skin lesion images acquired by means of epiluminescence surface microscopy and polarized-light videomicroscopy," Clin. Dermatol, vol. 20, no. 3, pp. 222–227, 2002.

[6] H. P. H Kittler, K Wolff, M Binder, "Diagnostic accuracy of dermoscopy," The Lancet Oncology, vol. 3, no. 3, pp. 159-165, 2022, doi: https://doi.org/10.1016/S1470-2045(02)00679-4.

[7] S. Y. Pan, "Q.: A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1345-1359, 2010.

[8] A. Esteva, Kuprel, B., Novoa, "Dermatologist-level classification of skin cancer with deep neural networks," Nature vol. 542, pp. 115–118 2017.

[9] J. S. U. Achim Hekler, Alexander H. Enk, Axel Hauschild, Michael Weichenthal, R.C. Maron, "Superior skin cancer classification by the combination of human and artificial intelligence," European Journal of Cancer, vol. 120, pp. 141-121, 2019, doi: https://doi.org/10.1016/j.ejca.2019.07.019.

[10] 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.

[11] C. G. T.Chen, "XGBoost: a scalable tree boosting system," arXiv:1603.02754, 2016.

[12] J. Zhang, Y. Xie, Q. Wu, and Y. Xia, "Skin lesion classification in dermoscopy images using synergic deep learning," in Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11, 2018: Springer, pp. 12-20.

[13] N. C. F. C. D. Gutman, E. Celebi, B. Helba, M. Marchetti, N. Mishra, A. Halpern, "Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) " arXiv:1605.01397, 2016.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," ed: Banff, Canada, 2013.

[15] I. J. Goodfellow et al., "Generative adversarial nets," Advances in neural information processing systems, vol. 27, 2014.

[16] J. T. Springenberg, "Unsupervised and semi-supervised learning with categorical generative adversarial networks," arXiv:1511.06390, 2015.

[17] S. A. C.Baur, N.Navab, "MelanoGANs: high resolution skin lesion synthesis with GANs,," arXiv:1804.04338, 2018.

[18] L. M. A. Radford, S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv:1511.06434, 2015.

[19] E. L. Denton, S. Chintala, and R. Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks," Advances in neural information processing systems, vol. 28, 2015.

[20] F. P. A.Bissoto, E.Valle,S.Avila, "Skin Lesion Synthesis with Generative Adversarial Networks," arXiv:1902.03253, 2019.

[21] A. K. (Owner). Skin Cancer ISIC. [Online]. Available: https://www.kaggle.com/datasets/nodoubttome/skin-cancer9-classesisic

[22] F. P. Alceu Bissoto, Eduardo Valle, Sandra Avila, "Skin Lesion Synthesis with Generative Adversarial Networks," arXiv:1902.03253, 2019, doi: https://doi.org/10.48550/arXiv.1902.03253.

[23] K. Djaroudib, P. Lorenz, R. Belkacem Bouzida, and H. Merzougui, "Skin Cancer Diagnosis Using VGG16 and Transfer Learning: Analyzing the Effects of Data Quality over Quantity on Model Efficiency," Applied Sciences, vol. 14, no. 17, p. 7447, 2024. [Online]. Available: https://www.mdpi.com/2076-3417/14/17/7447.

[24] H. Cheng, J. Lian, and W. Jiao, "Enhanced MobileNet for skin cancer image classification with fused spatial channel attention mechanism," Scientific Reports, vol. 14, no. 1, p. 28850, 2024/11/21 2024, doi: 10.1038/s41598-024-80087-w.

[25] K. S. K. B. Swathi, S. S. Chakravarthi, G. Ruthvik, J. Avanija and C. C. M. Reddy, "Skin Cancer Detection using VGG16, InceptionV3 and ResUNet," presented at the 4th International Conference on Electronics and Sustainable Communication Systems (ICESC), 2023.

A Hybrid AI Framework for DDoS Detection and Mitigation in SDN Environments Using CNN, GAN, and Semi-Supervised Learning

Abdelhakim HADJI, Brahim RAOUYANE
Department of Mathematics and Computer Science-Faculty of Sciences Ain Chock,
Computer Sciences and Systems Laboratory, Hassan II University of Casablanca, Casablanca, Morocco

***Abstract*—The fast technological evolution seen in recent years enhanced the performance and scalability of cloud computing infrastructure and Software-Defined Networking architectures. SDN provides programmability, centralized orchestration, and dynamic resource provisioning, while separating the control and data planes to offer promising architectural paradigm for cloud computing environments. Openness and flexibility expose SDN-based networks to other security concerns, such as large-scale Distributed Denial of Service (DDoS) attacks. This paper introduces a hybrid artificial intelligence (AI) framework for detecting and mitigating DDoS attacks in SDN environments. The framework leverages three complementary approaches: Convolutional Neural Networks (CNN) to capture temporal traffic patterns, Generative Adversarial Networks (GAN) to generate synthetic traffic for dataset augmentation and to enhance anomaly detection, and semi-supervised learning techniques to exploit large amounts of unlabeled traffic data. The proposed system is deployed on a testbed combining OpenDaylight as the SDN controller and Mininet for network emulation, while the AI models are trained and run in Anaconda environment. The network traffic flows are collected, processed into statistical features (i.e., packet rates, entropy values, protocol distribution ratios), and analyzed through the hybrid AI pipeline. Mitigation actions are configured through ODL RESTCONF interface, converting the detection into OpenFlow rules to drop or rate-limit the malicious packets. Experimental evaluation demonstrates that the proposed approach achieves high accuracy detection and robustness to unseen attacks patterns demonstrating the value of applying a hybrid CNN, GAN, Semi-supervised learning approach.**

***Keywords*—SDN; CNN; GAN; DDOS; OpenDaylight; Mininet; semi-supervised learning; hybrid AI framework**

I. Introduction

Over the past ten years, advances in technology significantly improved the performance and scalability of both cloud computing infrastructures and Software-Defined Networking (SDN) architectures. Caller in SDN architecture decouples the control plane from the data plane enabling, programmability, centralized orchestration and agile deployment of services to create a more efficient opportunity over a traditional network. This is the initial method to create a new innovative resource into Network Infrastructure as a key paradigm for the next-generation network that supports cloud-based services. However, the same level of openness and flexibility that provides efficiency and potential for innovation also introduces new vulnerabilities, Distributed Denial of Services (DDoS) attacks which remains one of the most significant security challenges to cloud and SDN/adoption of SDN infrastructures.

Conventional supervised machine learning techniques have achieved excellent performance on known DDoS patterns, but they are limited to known DDoS attacks due to the reliance on labeled datasets. [1]. Recent deliberations detailed in the literature indicate that hybrid models that combine supervised techniques with unsupervised and semi-supervised techniques will be favorable in settings with limited labelled data or high costs of labeling their data. Semi-supervised models have been a great way to take advantage of the ubiquitous unlabeled network traffic and exploit it to improve detection framework robustness and reduce annotation costs [2].

In this regard, Convolutional Neural Networks (CNNs) have emerged as an influential force in the realm of network security.

- Intrusion Detection: CNNs are adept at the detection of complex temporal and spatial patterns in traffic flows and thus can be effectively applied to intrusion detection in the environment of software defined networks (SDNs). For example, a CNN model was able to achieve an accuracy of 99.75% when detecting attacks using SDN–IoT networks [3].
- Feature Extraction: CNNs are capable of automatically extracting complex spatial features from high-dimensional traffic [4] data which is required to distinguish between legitimate and malicious flows [4].

In a similar manner, Generative Adversarial Networks (GANs) have also shown potential having significant ramifications for cybersecurity.

- Anomaly Detection: GANs can produce synthetic traffic that resembles legitimate traffic flows, therefore highlighting subtle deviations that may be indicative of zero-day attacks or progressive malware in cloud systems (L & Kousar, 2024).
- Adversarial Training: GANs produce synthetic adversarial DDoS traffic to increase the resilience of detection systems to advanced style attack strategies and adaptive adversaries [5].

While convolutional neural networks (CNNs) and generative adversarial networks (GANs) offer important advantages in terms of detection accuracy and robustness, the adaptation of these methods to changing threat vectors and the consistency of detection performance across heterogenous cloud-SDN environments presents additional challenges. Hybrid deep learning, adversarial modeling, and semi-supervised algorithms would address resilience, as well as detection accuracy and adaptivity.

Although the advancement of machine learning and deep learning applications to DDoS detection is increasing, the literature is still limited in ways in which the contributions of the literature can be invoked. A great deal of research relies on the use of fully labeled datasets, which diminishes their ability to generalize and adapt to attacks that are occurring for the first time. Many papers evaluate DDoS detection algorithms in some isolated or simulated environment and often do not deploy them in a controlled environment with a real SDN controller or in some emulated cloud environment. A great deal of research focuses on improving accuracy but does not consider adaptability and real-time mitigation mechanisms to handle DDoS attacks in SDN-enabled cloud environments. Thus, the importance of this research by offering a hybrid AI framework that combines CNNs, GAN, and Semi-supervised learning in an SDN-cloud testbed with OpendDaylight SDN controller and the Mininet emulator, which is capable of enhancing robustness, scalability, and adaptability within hybrid infrastructures where DDoS detection and mitigation can be improved.

In this paper, we present a hybrid AI framework to detect and mitigate distributed denial-of-service (DDoS) attacks in SDN-enabled cloud environments. The framework was implemented on an experimental testbed to represent an SDN-enabled cloud network, integrating OpenDaylight (ODL) as the control implementation and Mininet as the network emulation tool. The proposed method generates a traffic trace dataset that extracts and analyzes statistical features, such as packet rates, entropy measurements, and protocol ratios through the hybrid AI framework. Mitigation actions were dynamically enforced through ODL's RESTCONF API, acting on detection results and mapping them to OpenFlow rules to drop packets or set rates on egress traffic. Our results demonstrate that this hybrid integration enhances detection accuracy, decreases dependence on labeled datasets, and increases resilience against changing and customizing DDoS attack vectors.

II. Related Work

A. Machine Learning for SDN Security

The application of machine learning (ML) in securing Software-Defined Networking (SDN) structures from Distributed Denial of Service (DDoS) attacks has garnered growing attention in the research community. A recent study conducted by Musmuharam and Suharjito (2023) investigates the performance of multiple classification algorithms—specifically Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—alongside feature selection techniques such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), and t-Distributed Stochastic Neighbor Embedding (t-SNE). The authors found that the KNN model in conjunction with RFE produced highly significant results, achieving an accuracy of 99.97%, precision of 99.98%, recall of 99.96%, and F1 score of 99.97% for DDoS detection in an SDN context [6]. This indicates the effectiveness of custom feature engineering in improving the performance of traditional classifiers.

Complementary research [7] proposed a broader ML-based framework for cyber threat detection in SDN.This academic article examines the examination of cyber threats detection in Software-Defined Networks (SDNs) with a framework of machine learning methodologies by utilizing AI-Enabled Servers at the control and application layers to protect the network from them via identification and remediation of threats. The model makes use of six separate classifiers that are trained on differing metric parameters in the network: Random Forest; K-Nearest Neighbor; Support Vector Machine; Naïve Bayes; Decision Tree; and Logistic Regression, whereby analysis of the data supports that the framework is capable of detecting and remediating risks of a cyber nature; and also proposes recommendations for more effective threat detection in the future, and future security controls.

B. CNNs in Network Security

Convolutional Neural Networks (CNNs) have demonstrated utility for intrusion detection and anomaly detection of systems with Software Defined Networking (SDN) properties, primarily due to their ability to automate spatial and temporal feature extraction from challenging data related to network traffic.

Others examined the weakness of CNNs as a means for developing Intrusion Detection Systems (IDS) for SDNs with respect to the overfitting phenomenon that often degrades the performance of deep learning development[8]. By implementing two new approaches based on regularization or uncertainty with their new learning method of CNNs, they demonstrated their improved resilience to unseen attack. Training and testing the learning method on the InSDN benchmark dataset generated higher generalization for the model and better detection accuracy for unknown intrusion events to enhance security for SDN infrastructures against attacks.

In a more recent study, and with a focus on lightweight DDoS detection and mitigation for SDN, [9] a security framework that integrated an anomaly detection and prevention module utilizing a hybrid CNN-MLP model has been proposed. The design leveraged the ability of CNNs to perform feature extraction with the efficiency of multilayer perceptron's (MLPs) since it provided real-time analysis of network traffic; therefore, their system had the ability to identify anomalies, and respond using flow-rule instructions, premised in the SDN controller, to block attacker IP addresses and mitigate malicious traffic. Overall, they were able to demonstrate cost-effective and real-time solutions, thus indicating the feasibility of hybrid CNN architectures to protect against expensive resource DDoS attacks in SDN time-critical infrastructure.

Collectively, these works demonstrate that CNNs play an important role that can significantly bolster security properties for SDN. They further demonstrate that CNNs bearing the

capability of flexibility and when regularized and/or hybridized with other levels of neural architectures, bring robustness and efficiency to the process of identifying and mitigating many classes of cyber threats.

C. GANs in Cybersecurity

Generative Adversarial Networks (GANs) are now increasing attention within cybersecurity, especially in relation to mitigating the limitations of SDN-based infrastructures. Their two-fold potential of generating realistic synthetic traffic and supporting anomalous detection have made GANs valuable in both research and operational defense areas.

A motivating example of GANs used to support SDN-GAN security was presented [10] to introduced one of the early applications of SDN-GAN, a framework that synthesizes cyber-attack scenarios. They illustrated a way that GANs could autonomously synthesize realistic attack signatures based on known patterns. Doing so enables the training of the detection systems in the endless number of possible attack variations. In this regard, the challenge of SDN security is one that is most vulnerable to programmable environments with frequently unknown and adaptive threats.

Another study has recently introduced the LD-BiHGA system [11], a bi-channel hybrid GAN architecture designed for anomaly detection in multi-domain SDN systems. In their model, the two asymmetric GANs were used for independent learning of normal and abnormal traffic, while the bi-channel attention mechanism provided an additional level of feature extraction. By overcoming concerns of real-time, unbalanced traffic data, the LD-BiHGA system improved accuracy in recognition of anomalies over traditional intrusion detection systems, demonstrating the effectiveness of GANs utilized in hybrid models for larger-scale, complex SDN systems. Overall, these contributions signify the journey of GANs in the cyber security space; from synthetically generating attack scenarios, hybridizing with RNNs, and attention mechanisms for anomaly detection. They exemplify not only the capacity for dataset enhancement and classification imbalance for GANs, but also the viability for adaptive engagement and real-time security within SDN infrastructures.

D. Semi-Supervised Learning for SDN Security

In recent years, semi-supervised learning has gained considerable attention as a viable strategy for detecting DDoS attacks in SDN and cloud-based environments, mainly due to the algorithm's capability of utilizing large quantities of unlabeled traffic data. In contrast to purely supervised techniques, which typically depend on having access to large labeled datasets, semi-supervised learning proves a more practical and viable option for real-world implementations.

In addition, a survey of machine learning techniques for intrusion detection in cloud and SDN environments [12]. They found that ML models can detect anomalous traffic related to DDoS attacks and referenced a clear shift towards adaptive rule-based models to increase the security of routing payloads while preserving network privacy.

In IoT-SDN environments, Ravi and Shalinie (2020) introduced the LEDEM mechanism, which acts as a learning-executed detection and mitigation framework. The LEDEM mechanism demonstrated an accuracy of 96.28% for DDoS detection and established the combination of cloud and SDN architecture with learning techniques as successful[13]. This research shows how dynamic and distributive detection models can benefit resource-constrained infrastructures relying on IoT-SDN.

Khamaiseh et al. [2]examined the effectiveness of supervised versus semi-supervised methods for detecting both known and unknown attacks. Their results indicated that supervised methods performed well when encountering previous attack patterns, however, performance suffered against zero-day attacks, demonstrating the limit of these methods. In comparison, semi-supervised methods demonstrated improved adaptability and resilience, and both authors argued for their eventual role in securing DDoS attacks, especially as part of a future SDN security framework. Deka et al. offered a dynamic, range-based anomaly detection method leveraging semi-supervised algorithms[14]. Their work involved the introduction of a density-based approach to handling unlabeled traffic. To that end, semi-supervised classifiers outperformed other standard supervised classifiers in terms of scalability and accuracy offered in fast changing SDN situations.

Overall, these studies contribute to a growing body of studies acknowledging the value of semi-supervised learning method for reducing the number of labeled data needed to detect attacks regardless of their pattern, and combining cost-effective label data and semi-supervised methods will lead to shared improvements to future DDoS detection and security frameworks.

The literature reviewed in this research clearly indicates that there have been significant developments in the detection and prevention of DDoS attacks, especially in SDN and cloud-based contexts, through machine learning (ML) and deep learning (DL) methods. Classical ML approaches, e.g., Random Forest, Support Vector Machine, and K-Nearest Neighbors, achieved relatively high detection accuracy levels when employed with optimally selected features. However, these ML models remain limited by their reliance on labeled datasets and their limited ability to generalize to new or even zero-day attack datasets.

Deep learning methods, namely convolutional long short-term memory (CNN-LSTM), have shown strong performance in automatically extracting important spatial-temporal features from complex traffic data. Regularization methods and hybrid CNN approaches have also shown the ability to improve generalization performance and reduce overfitting levels, which has also made them more robust for designing an intrusion detection task. In a similar vein, GANs have opened important new lines of research into anomaly detection both in isolation and in conjunction with adversarial training. Additionally, GANs can be employed to generate synthetic traffic data that help balance datasets and create realistic attack data that improves performance robustness. Despite these advances, adapting GAN models to solve multi-domain and real-time SDN settings and cloud contexts still presents significant challenges in computational overhead and interpretability.

Semi-supervised learning has emerged as an exciting alternative to address one of the central challenges of SDN security: the limited/missing labeled data. Previously, it has been shown that semi-supervised algorithms can leverage unlabeled traffic due to its abundance, allowing for higher adaptability to altered attacks without increased annotation costs. While the performance of these methods is promising, their placement back into operational environments is still sparse, particularly in dynamic enforcement and scalability.

In summary, the current state-of-the-art shows clear advancements, while identifying areas for research. Most work employs supervised/deep learning models (CNN, GAN) or simply employs semi-supervised anomaly detection. Very few papers propose a holistic framework that employs CNNs for deeper feature learning, GANs for synthetic traffic generation and adversarial accuracy, and a semi-supervised model for changes to new attacks. Additionally, few papers demonstrate end-to-end pipelines that not only detect attacks but also capable of real-time mitigation via interacting directly with SDN Controllers such as OpenDaylight.

This outlines the issue of missing a unified model; the present work proposes a hybrid AI framework of CNN, GAN and semi-supervised algorithms that can create a real-time effective DDoS detection and mitigation framework for SDN based cloud environments.

III. Methodology

A. General Architecture

The architecture of the proposed framework integrates concepts of Software-Defined Networking (SDN) and Artificial Intelligence (AI) paradigms to automate the process of detection, mitigation, and protection against Distributed Denial of Service (DDoS) attacks. Two virtual machines compose the experimental testbed set-up, one contains an OpenDaylight (ODL) controller that will provide centralized control and policy enforcement, while the other contains the network emulator Mininet, which creates the data plane for the framework via an Open vSwitch (OVS) and manufactures both legitimacy and harmful traffic. A host machine running Jupyter/Anaconda will provide the environment for AI processing, wherein detection models are established and applied on this machine.

The three layers of architecture follow the principles of a SDN:

- Application Layer: AI driven detection applications (CNN, GAN, semi-supervised algorithms).
- Control Layer: ODL controller with RESTCONF APIs to enforce policies based on the detection of malicious traffic.
- Data Layer: Mininet switches and hosts that contain the programmable data plane.

B. Data Collection and Feature Engineering

Traffic is generated within Mininet, consisting of benign traffic flows (HTTP, FTP, DNS) and malicious traffic flows (SYN floods, UDP floods, ICMP floods). It is captured using tcpdump/tshark and exported into .pcap files then transformed into CSV datasets (traffic.csv).

Statistical and time-based features are extracted from each time window of traffic including:

- Packet-level: packets per second (pps), bytes per second (bps), average packet size.
- Protocol distribution: SYN/ACK ratio, UDP/TCP ratio, ICMP fraction.
- Diversity of sources: number of unique source IPs and the entropy of the source IP distribution.
- Time-based features: inter-arrival time variance, coefficient of variation.

These provided the inputs to the hybrid AI models.

C. Integration of AI Models

The main contribution of our architecture is the hybrid AI pipeline that utilizes the combination of Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and semi-supervised learning.

1) Convolutional Neural Networks (CNNs): CNNs will be utilized to capture rich temporal-spatial dependencies across traffic features. CNNs hierarchical architecture allows automatic feature extraction and effective classification of DDoS versus benign flows.

2) Generative Adversarial Networks (GANs): GANs will be used in two ways: 1) for augmenting the dataset by generating realistic benign/malicious traffic to compensate for class imbalance, and 2) for adversarial training, where the generator produces difficult samples, thereby increasing detection accuracy of the discriminator.

3) Semi-supervised learning: Real-world datasets often do not have enough labeled samples, so semi-supervised methods (e.g., Self-Training Classifiers, consistency regularization) will be used to utilize unlabeled traffic, improving adaptability to emerging or zero-day attacks.

The final outputs from the CNNs (supervised predictions), the GAN discriminators (anomaly detection signal), and the semi-supervised classifiers will all go through a late-fusion mechanism producing an adapted detection score that optimizes accuracy and adaptability.

D. Policy Engine and Mitigation

Upon detection and classification of a given traffic flow, the decision is passed through a RESTCONF northbound API to the ODL controller. The ODL controller implements the decision by converting the identified flows into OpenFlow rules and implementing those rules onto the Mininet switches. Mitigation includes:

1) Flow-based filtering: implement DROP rules based on the identified malicious source IP or port.

2) Rate-limit (meter tables): impose a maximum bandwidth constraint on a suspected malicious flow.

3) Dynamic updates: use short flow timeouts to eventually rollback flows in case they are a false positive.

The combination of AI detection across the traffic flows and the programmability of an SDN approach will ultimately enable real-time, dynamic mitigation against a DDoS attack.

Algorithm 1: Hybrid AI Framework for DDoS Detection and Mitigation in SDN

```
Initialize: SDN environment
Compute: extract statistical features from captured network traffic in Mininet (normal and abnormal)
While (new traffic window is received) do
    For (each traffic flow in the window) do
    Update feature vectors for the flow
    Analyze flow using AI Models:
        If (fusion score ⩾ threshold) then
        Search for malicious flows
        Update ODL via RESTCONF:
            If (attacked) then
                Insert OpenFlow DROP rule
                Or apply METER for rate-limiting
    End
End
```

As a summary, the framework can be thought of as a pipeline:

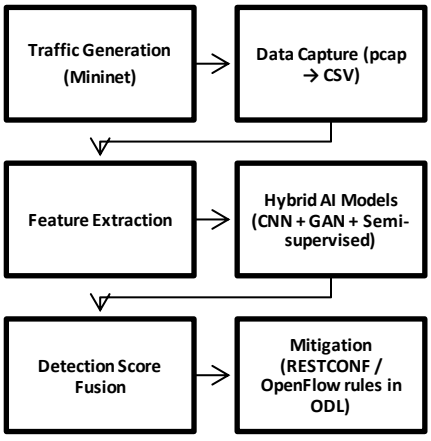


Fig. 1. Framework pipeline.

This design guarantees that detection is accurate (with CNN), resilient (with GAN), and adaptive (through semi-supervised learning), while mitigation is performed immediately in the SDN data plane. Fig. 1 shows framework pipeline.

IV. Results and Discussion

The experimental evaluation of this project was built with Mininet (data plane) and OpenDaylight (control plane), with the application layer running in a Jupyter/Anaconda working environment for the AI models developed in Python. Traffic patterns included normal HTTP, FTP, and DNS flows, as well as malicious traffic created using tools like hping3 to simulate SYN floods, UDP floods, and ICMP floods.

TABLE I. Results in a Controlled Environment

Model	Evaluation Metrics			
	Accuracy (%)	*Precision (%)*	*Recall (%)*	*F1-score (%)*
CNN	99.0	98.5	98.0	99.0
GAN	97.5	97.0	96.5	96.8
Semi Supervised	95.0	94.0	95.0	95.2

The CNN model was able to classify all normal traffic, as well as known DDoS attack types, with high accuracy, reaching a maximum accuracy/F1-score of just below 99% when trained on balanced datasets as shown in Table I. However, the CNN model degraded in accuracy when faced with new attack types that had not been included in the training data, which is a limitation of any supervised model.

The GAN model performed even more robustly by generating synthetic benign and DDoS samples, improving generalizability across imbalance datasets that had far fewer malicious flows than benign flows. Additionally, the GAN discriminator component performed well as an anomaly detector for all types of network traffic, achieving high accuracy in classifying flows as benign or malicious based on their deviation from normal distribution.

The semi-supervised approach made significant gains when access to labeled data was limited. The model was able to take advantage of pseudo-labeling and consistency regularization, which allowed it to utilize unlabeled flows during the training process, rather than depending on labeled annotations. When only 30 percent of the dataset was labeled, the semi-supervised structure achieved F1-scores higher than 95 percent, which surpassed traditional supervised baselines with the same labeling restrictions.

Although results obtained from single models demonstrate promise, weaknesses in stand-alone models justify hybridization. The proposed framework from this study uses late fusion by combining CNN, GAN, and Semi-supervised learning method's output through late fission logic. In late fusion, the output of the supervised probabilities is combined with the anomaly scores as a composite decision.

Table II presents the performance improvement observed before/after using a fusion approach.

The use of a fusion approach reduced the false positives over the previous stand-alone models. In instances of zero-day events (e.g. new UDP amplification traffic), the base model hybrid captures 95% recall whereas the CNN model captured 88%.

TABLE II. DETECTION PERFORMANCE BEFORE AND AFTER HYBRID FUSION

Model	Accuracy	Recall (%)	False Positive Rate (%)
CNN	99.0	88.0	6.5
GAN	97.5	90.0	7.2
Semi Supervised	95.0	92.0	6.8
hybride	98.7	95.0	4.3

Detection results were directly incorporated into the ODL controller via the RESTCONF northbound interface. Once malicious flows were identified, ODL automatically pushed OpenFlow rules into Mininet switches for immediate execution. The following two countermeasures were evaluated Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.

V. CONCLUSION

The findings validate that the suggested framework can balance accuracy, adaptability, and operational feasibility. CNNs can recognize structured patterns, GANs allow for synthetic augmentation, and adversarial training further enhances robustness against threat behavior as it was. Models as we have indicated in the semi-supervised methods augment adaptability to unknown threats. We also find that the hybrid implementation of these models in an SDN testbed demonstrates the practical feasibility of AI-enabled hybrid detection frameworks for today's emerging, programmable networks.

Despite this success, there are still challenges to overcome. GAN training on its own has a heavy computational burden which, all other factors constant, may limit real-time scalability of the solution when deployed in large-scale integrations. The semi-supervised models, while still viable, are sensitive to precise tuning of the pseudo-labeling threshold to prevent erroneous error propagation in known threat behavior in the etiologies projected.

Future improvements in our integration should consider a distributed learning approach (e.g., federated learning) and/or a model optimization (e.g., pruning) to deploy when computational resources are confined to the edge environments.

REFERENCES

[1] V. Vivek and B. Veeravalli, "A Survey on Machine Learning Approaches for Intrusion Detection in Cloud Computing Environments for Improving Routing Payload Security and Network Privacy," in 2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), July 2024, pp. 79–85. doi: 10.1109/IAICT62357.2024.10617793.

[2] S. Khamaiseh, A. Al-Alaj, M. Adnan, and H. W. Alomari, "The Robustness of Detecting Known and Unknown DDoS Saturation Attacks in SDN via the Integration of Supervised and Semi-Supervised Classifiers," Future Internet, vol. 14, no. 6, p. 164, June 2022, doi: 10.3390/fi14060164.

[3] V. Sujatha and S. Prabakeran, "Lightweight DDoS Attack Detection and Mitigation in Software-Defined Networks Using Deep Learning," in 2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS), Oct. 2023, pp. 1–8. doi: 10.1109/ICCAMS60113.2023.10525696.

[4] K. S. Goud and G. S. Rao, "Towards an Efficient DDoS Attack Detection in SDN: An Approach with CNN-GRU Fusion," in 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), Jan. 2024, pp. 1–10. doi: 10.1109/ICAECT60202.2024.10469528.

[5] "Generative Adversarial Networks for Malware Detection in Cloud Computing Environments | Request PDF," ResearchGate, Aug. 2025, doi: 10.26562/ijiris.2024.v1004.24.

[6] "Detection of Distributed Denial of Service Attacks in Software Defined Networks by Using Machine Learning," Int. J. Commun. Netw. Inf. Secur., Nov. 2023, doi: 10.17762/ijcnis.v15i3.6214.

[7] I. Sharma, J. Saini, G. Chhabra, and K. Kaushik, "Cyber Threat Detection in Software-Defined Networks: An Empirical Analysis of Machine Learning Methods," in 2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Dec. 2023, pp. 1119–1124. doi: 10.1109/ICIMIA60377.2023.10426240.

[8] "(PDF) The role of CNN for Intrusion Detection Systems: An Improved CNN Learning Approach for SDNs," in ResearchGate, Accessed: Sept. 25, 2025. [Online]. Available: https://www.researchgate.net/publication/351779322_The_role_of_CNN_for_Intrusion_Detection_Systems_An_Improved_CNN_Learning_Approach_for_SDNs

[9] V. Sujatha and S. Prabakeran, "Lightweight DDoS Attack Detection and Mitigation in Software-Defined Networks Using Deep Learning," in 2023 International Conference on New Frontiers in Communication, Automation, Management and Security (ICCAMS), Oct. 2023, pp. 1–8. doi: 10.1109/ICCAMS60113.2023.10525696.

[10] A. AlEroud and G. Karabatis, "SDN-GAN: Generative Adversarial Deep NNs for Synthesizing Cyber Attacks on Software Defined Networks," in On the Move to Meaningful Internet Systems: OTM 2019 Workshops, C. Debruyne, H. Panetto, W. Guédria, P. Bollen, I. Ciuciu, G. Karabatis, and R. Meersman, Eds., Cham: Springer International Publishing, 2020, pp. 211–220. doi: 10.1007/978-3-030-40907-4_23.

[11] S. Prabu and J. Padmanabhan, "Bi-channel hybrid GAN attention based anomaly detection system for multi-domain SDN environment," J. Intell. Fuzzy Syst., vol. 46, no. 1, pp. 457–478, Jan. 2024, doi: 10.3233/JIFS-233668.

[12] V. Vivek and B. Veeravalli, "A Survey on Machine Learning Approaches for Intrusion Detection in Cloud Computing Environments for Improving Routing Payload Security and Network Privacy," in 2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), July 2024, pp. 79–85. doi: 10.1109/IAICT62357.2024.10617793.

[13] N. Ravi and S. M. Shalinie, "Learning-Driven Detection and Mitigation of DDoS Attack in IoT via SDN-Cloud Architecture," IEEE Internet Things J., vol. 7, no. 4, pp. 3559–3570, Apr. 2020, doi: 10.1109/JIOT.2020.2973176.

[14] P. Kr. Deka, Y. Verma, A. B. Bhutto, E. Elmroth, and M. Bhuyan, "Semi-Supervised Range-Based Anomaly Detection for Cloud Systems," IEEE Trans. Netw. Serv. Manag., vol. 20, no. 2, pp. 1290–1304, June 2023, doi: 10.1109/TNSM.2022.3225753.

Benchmarking Deep Learning Models for Visual Classification and Segmentation of Horticultural Commodities

Fuzy Yustika Manik*, Syahril Efendi, Jos Timanta Tarigan, Maya Silvi Lydia
Department of Computer Science-Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

***Abstract*—Recent advances in computer vision have enabled new approaches for automated quality assessment of tropical fruits, where accurate classification and segmentation are essential for postharvest inspection. A major challenge lies in identifying deep learning architectures that achieve high accuracy while remaining computationally efficient for potential edge-based deployment. This study benchmarks three Convolutional Neural Network (CNN) models for classification (VGG16, ResNet50, and EfficientNet-B0) and two encoder–decoder models for segmentation (U-Net and DeepLabV3+) using annotated pineapple and strawberry image datasets. A 5-fold cross-validation strategy was applied to ensure statistical robustness, with evaluation metrics including accuracy, precision, recall, F1-score, Intersection over Union (IoU), and Dice coefficient. Statistical significance was verified using the Friedman and Wilcoxon signed-rank tests ($\alpha = 0.05$ and 0.01). EfficientNet-B0 achieved the best classification results with average accuracies of 91.4% (strawberry) and 90.7% (pineapple), significantly outperforming ResNet50 and VGG16 ($p < 0.01$). For segmentation, DeepLabV3+ obtained the highest performance with mean IoU values of 91.7% and 90.8% and Dice coefficients above 92%, indicating precise boundary delineation of ripe and defective regions. Computational efficiency analysis further showed that EfficientNet-B0 had the lowest inference time (0.026 s) and smallest model size (20.4 MB), making it ideal for real-time or embedded applications. Visual analysis confirmed that DeepLabV3+ maintained robustness at fruit boundaries, though minor misclassifications were observed. This benchmarking highlights the combination of EfficientNet-B0 and DeepLabV3+ as a reliable baseline for deep learning-based fruit quality assessment.**

***Keywords*—Fruit quality assessment; classification; segmentation; EfficientNet-B0; DeepLabV3+; AISAM-CSNet**

I. Introduction

Advances in computer vision and deep learning have significantly impacted precision agriculture, particularly in automated fruit quality assessment and post-harvest monitoring [1], [2]. One of the most crucial visual tasks in this domain is image-based classification and segmentation, which enables sorting, grading, and intelligent decision-making for horticultural commodities such as strawberries and pineapples. These visual tasks are particularly challenging due to variations in lighting, occlusion, background complexity, and the diversity of shapes and colors within a single fruit class [3].

The application of deep learning in precision agriculture systems has received significant attention in recent years. In particular, Convolutional Neural Network (CNN)-based models such as VGG16, ResNet50, and EfficientNet-B0 have demonstrated high performance in fruit image classification tasks, ranging from type determination and ripeness to visual defect detection [4], [5].

A study by Hasan et al. [6] demonstrated that ResNet50 can accurately classify fruit in real-world scenarios. Meanwhile, VGG16 is known to be a stable model and is often used as a baseline in many visual classification experiments [7]. EfficientNet-B0, with its systematic scaling approach, offers a balance between accuracy and computational efficiency, making it suitable for edge-based applications such as field devices [8]. However, most of these models are trained and tested under relatively controlled conditions or using datasets with low levels of variation. In real-world horticultural scenarios, fruit images are typically acquired under non-uniform lighting, varying backgrounds, and inconsistent fruit shapes and sizes [8].

U-Net has become a primary choice for fruit image segmentation tasks due to its symmetric encoder-decoder design, which effectively preserves spatial information [9]. On the other hand, DeepLabV3+ uses atrous convolution and atrous spatial pyramid pooling (ASPP) to capture contextual information at multiple scales, proving effective in segmenting objects in complex backgrounds [10], [11]. Mo et al. [12] successfully implemented MobileNetV2-based DeepLabV3+ to detect sugar apple ripeness accurately.

However, most of these models were trained and tested under relatively controlled conditions or using datasets with low levels of variation. In real-world horticultural scenarios, fruit images are typically acquired with non-uniform lighting, varying backgrounds, and inconsistent fruit shape and size [13]. These factors reduce the generalization ability of standard CNN models. Therefore, several approaches such as data augmentation [14], domain adaptation [15], and synthetic training [16] have been developed to improve the generalization ability of CNN models to real-world conditions.

Furthermore, most previous studies focus on a single task (classification or segmentation) and often use datasets under ideal conditions. Few studies evaluate multiple CNN architectures simultaneously on both tasks on authentic tropical fruit images such as strawberries and pineapples. Therefore,

*Corresponding author.

this study aims to systematically assess the performance of VGG16, ResNet50, and EfficientNet-B0 models for classification and U-Net and DeepLabV3+ for tropical fruit image segmentation. The evaluation uses metrics such as accuracy, precision, recall, F1-score (classification), and IoU and Dice coefficient (segmentation) to provide a scientific basis for selecting the optimal model for image-based fruit classification and segmentation systems.

This benchmarking study aims to highlight the strengths and weaknesses of current deep learning approaches and provide an empirical basis for developing more adaptive and multi-task architectures. In addition, the findings of this study will demonstrate how model selection and architectural complexity directly influence accuracy, segmentation quality, and computational efficiency under real-world postharvest conditions. These insights contribute to the development of practical, lightweight, and adaptive deep learning models that can be effectively applied to agricultural automation and quality monitoring of tropical fruits in real-world scenarios. Based on our results, we also briefly discuss the potential of emerging immune-inspired and multi-agent approaches such as AISAM-CSNet, which will be elaborated in detail in our forthcoming publication.

II. Related Work

Rapid computer vision and deep learning developments have driven the application of CNN models in various sectors, including precision agriculture. In fruit image processing, two main tasks, quality classification and fruit object segmentation, are key for automated post-harvest systems. Recent studies have utilized modern CNN models to detect fruit types and ripeness and to separate fruit objects from complex backgrounds. CNN models such as VGG16, ResNet50, and EfficientNet-B0 are widely used in fruit image classification. VGG16 is a classic architecture often used as a baseline due to its stability despite the large number of parameters [7]. Sudars et al. [6] conducted a comprehensive review of the application of CNNs to fruit quality classification and positioned VGG16 as one of the standard architectures used in laboratory scenarios.

ResNet50, which introduces residual learning, effectively addresses degradation issues in deep networks [3]. Hasan et al. [6] showed that ResNet50 can maintain fruit classification accuracy in natural lighting. Arif et al. [17] compared ResNet50 with DenseNet and EfficientNet in orange classification, with ResNet50 outperforming in complex background conditions.

EfficientNet-B0 introduces a compound scaling approach to balance accuracy and efficiency [8]. Li et al. [5] demonstrated that EfficientNet-B0 suits fruit classification on edge devices. Wagle et al. [18]'s research corroborates this finding by showing the high computational efficiency of EfficientNet in apple and tomato classification. Reyes et al. [19] also demonstrated that EfficientNet is effective in fine-grained classification of tropical fruits. Furthermore, DenseNet [20], InceptionV3, and MobileNetV2 [21] were also evaluated in fruit classification, but their performance was often lower in open field conditions. A study by Rauf et al. [22] confirmed that ResNet and EfficientNet provide the best trade-off between accuracy and inference time compared to other architectures.

Meanwhile, fruit object segmentation requires precise object boundary detection and separation of the fruit from the background. The U-Net model is widely recognized for this task because it preserves spatial information [9]. Fang et al. [23] developed a U-Net with attention gating for accurate strawberry segmentation in open fields. Jamil et al. [24] combined a squeeze-and-excitation block to improve mango fruit detection in RGB images.

DeepLabV3+ utilizes atrous convolution and atrous spatial pyramid pooling (ASPP) to process multiscale features. Research by Zhao et al. [10] demonstrated the superiority of DeepLabV3+ in detecting mangoes under occlusion. Mo et al. [12] implemented MobileNetV2-based DeepLabV3+ to detect the ripeness of sugar apples and achieved high accuracy. Other models such as Mask R-CNN [25], HRNet [26], and SegNet [27] have also been tried for fruit segmentation, but their complexity and high computational requirements are obstacles in real-time applications. Milioto et al. [28] developed real-time semantic segmentation for crops and weeds using a CNN optimized for agricultural robots.

One of the main challenges in implementing CNN in fruit classification and segmentation systems is its ability to generalize real-world images. Most studies use datasets with clean backgrounds and ideal lighting [29]. Xu et al. [30] reported a 30% decrease in accuracy when apple classification models were trained in the laboratory and tested in the open field. To address this issue, data augmentation [14], domain adaptation [15], and synthetic image-based training [16] approaches have been proposed. Shorten and Khoshgoftaar [14] showed that augmentation can improve the robustness of CNN models to complex background conditions. Chen et al. [15] evaluated domain adaptation to transfer models from laboratory to field data. Meanwhile, synthetic training has enriched the variety of training images, as evidenced by Rahnemoonfar and Sheppard [16].

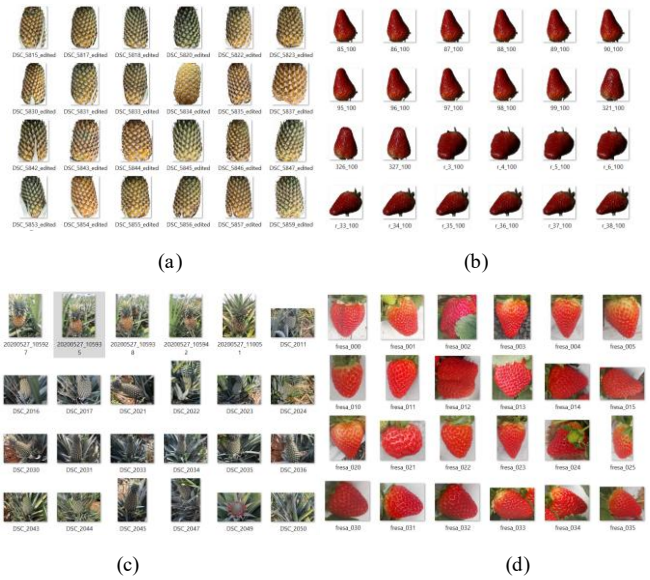
Although various CNN models have proven effective for fruit classification and segmentation separately, comprehensive studies that evaluate multiple CNN architectures across both tasks under real-world conditions remain limited. This research presents the first benchmarking study that simultaneously evaluates CNN-based classification and segmentation models (VGG16, ResNet50, EfficientNet-B0, U-Net, and DeepLabV3+) on tropical fruit images (strawberries and pineapples) in realistic postharvest scenarios. The study contributes a novel understanding of the relationship between architectural complexity, adaptability, and segmentation precision. The benchmarking results reveal a balanced trade-off between classification and segmentation accuracy as well as computational efficiency, forming the foundation for the development of AISAM-CSNet as a lightweight, adaptive, and multitask model for agricultural automation.

III. Methodology

The research stages include data acquisition and preprocessing, CNN model architecture development (classification and segmentation), implementation, and performance evaluation.

A. Dataset and Data Acquisition

This study uses strawberry images for non-climacteric fruits and pineapple images for climacteric fruits obtained from two main sources: 1) primary: a collection of field images obtained directly using a digital camera, and 2) secondary: public datasets such as Kaggel [31-32]. The image dataset was collected in real-world environments with natural variations in lighting and diverse surface textures. Manual annotation was performed based on the local postharvest conditions of tropical horticultural products. Examples of the images used can be seen in Fig. 1 and Fig. 2 below:



(a) (b) (c) (d)

Fig. 1. Secondary dataset: (a) Pineapple with white background, (b) Strawberry with white background, (c) Pineapple with natural background, (d) Strawberry with natural background.

(a) (b) (c) (d)

Fig. 2. Secondary dataset: (a) Strawberry with white background, (b) Strawberry with natural background, (c) Pineapple with white background, (d) Pineapple with natural background.

The distribution of the dataset based on quality categories and background types can be seen in Table I:

TABLE I. TYPE STYLES

Fruit	Category	Primary		Secondary		Initial Total
		White	Natural	White	Natural	
Strawberry	Good / Ripe	20	15	50	30	70
	Medium / Unripe	30	20	60	30	90
	Damaged / Rotten	15	10	20	15	35
Pineapple	Good / Ripe	15	10	50	20	65
	Medium / Unripe	20	15	60	30	80
	Damaged / Rotten	10	5	25	15	35

Note: "Primary" and "secondary" refer to the sources of the images in the dataset. Primary data is data captured directly using a digital camera, while secondary data is public datasets such as Kaggel data. The background types are either white or natural (real-world scenes). The "initial total" column represents the total number of images per quality category before preprocessing or augmentation.

The dataset is categorized into two subsets: Classification: Images labeled according to quality classes (good, medium, poor) or ripeness level, and Segmentation: Images annotated with fruit masks for each individual fruit object.

B. Preprocessing

To guarantee consistency and improve model resilience, every image was subjected to a uniform preprocessing pipeline before the training phase. Among the steps were:

- Resizing: To comply with the input specifications of the corresponding models, images were shrunk to 224×224 pixels for classification tasks and 256×256 pixels for segmentation tasks.
- Normalization: To enable quicker convergence during training, pixel values were scaled from the initial range of [0,255] to a normalized range of [0,1].
- Data Augmentation: Random augmentations such as ±10% zoom, ±20° rotation, and horizontal flipping were used to enhance the training dataset and enhance generalization.

Examples of the augmented photos for the pineapple and strawberry samples are shown in Fig. 3.

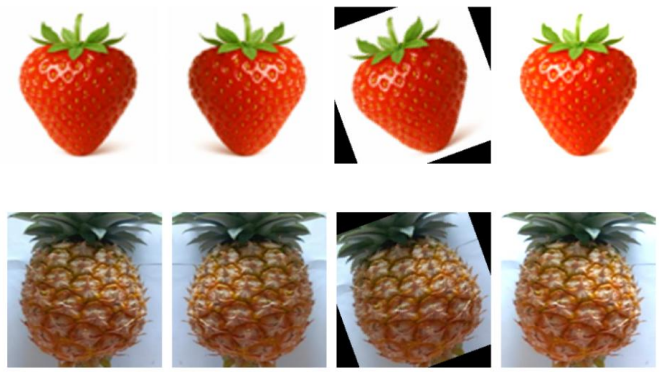


Fig. 3. Data augmentation examples for strawberry (top) and pineapple (bottom) images used to enhance model generalization.

C. Model Architectures Used

Five CNN architectures were selected as baselines for benchmarking classification and segmentation tasks.

1) VGG16: VGG-16 consists of 13 convolutional layers grouped into 5 blocks. Each block contains 2 or 3 convolutional layers. Every convolutional layer uses a 3×3 kernel, 'same'

padding, and the ReLU activation function. After each convolutional block, a 2×2 Max Pooling layer is applied to reduce the spatial dimensions (downsampling), preserve important features, and lower computational complexity. Once all convolutional and pooling blocks are completed, the extracted features are flattened and passed through 3 dense (fully connected) layers. The first two dense layers use ReLU activation, while the final dense layer uses a softmax activation for classification (see Fig. 4).

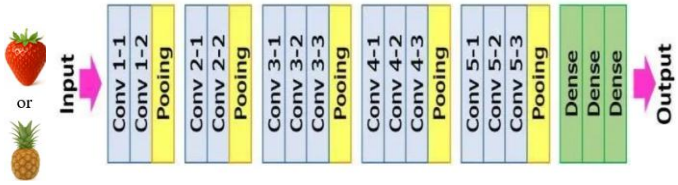


Fig. 4. Architecture of VGG 16.

2) EfficientNet: The architecture of EfficientNet, which consists of seven primary blocks intended for effective feature extraction from input photos, is depicted in Fig. 5 below. After a normal 3x3 convolution layer, a sequence of Mobile Inverted Bottleneck Convolutions (MBConv) is performed. While Blocks 3 to 6 use MBConv6 layers with 5×5 kernels to increase the receptive field without appreciably increasing the number of parameters, Blocks 1 and 2 use MBConv layers with 3×3 kernels. The extracted features are refined in the final Block 7 using an MBConv6 with a 3×3 kernel. High computational efficiency is made possible by the expansion, depthwise convolution, and projection processes included in each MBConv layer. The resulting feature map at the end of the network serves as a rich representation of the input, suitable for downstream tasks such as fruit quality classification. EfficientNet is particularly advantageous for image-based classification of fruits like strawberries and pineapples due to its balance between performance and resource efficiency.

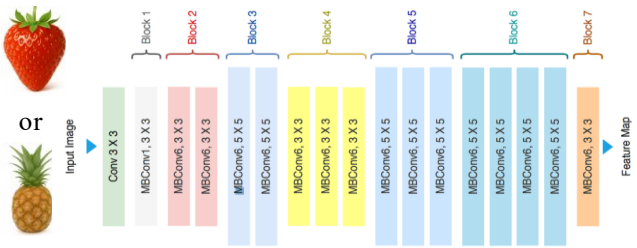


Fig. 5. Architecture of EfficientNet.

3) ResNet50: Fig. 6 given below illustrates the architecture of the ResNet50 model. The architecture begins with a zero padding process, followed by an initial convolutional layer composed of convolution, batch normalization, ReLU activation, and max pooling—collectively referred to as Stage 1. Next, the network comprises four main stages (Stage 2 to Stage 5), each consisting of a Conv Block and several Identity Blocks (ID Blocks). After passing through all the convolutional blocks, the network concludes with an average pooling layer, a flattening process, and a fully connected (FC) layer that produces the final class prediction output.

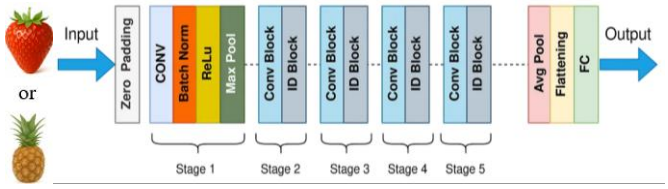


Fig. 6. Architecture ResNet50.

4) U-Net: The U-Net convolutional neural network architecture is specifically designed for image segmentation tasks, such as identifying decayed or damaged areas on pineapple and strawberry fruits. U-Net has a symmetrical structure resembling the letter "U" and consists of two main parts: the contracting path (encoder) on the left side and the expansive path (decoder) on the right. It processes input images of size 256×256×3 (RGB) through the encoder path, which includes 3×3 convolutional blocks with ReLU activation followed by 2×2 max pooling operations. This progressively increases the number of feature channels (from 32 to 512) while reducing the spatial dimensions. At the deepest part of the network (the bottleneck), complex features are represented at the smallest resolution (8×8) with a depth of 512 channels. The decoder then reconstructs the spatial dimensions using 2×2 up-convolution operations, while skip connections from the encoder help retain spatial details. The process concludes with a 1×1 convolution that outputs a segmented image of size 256×256×3, precisely identifying decayed and non-decayed areas on the fruit (see Fig. 7).

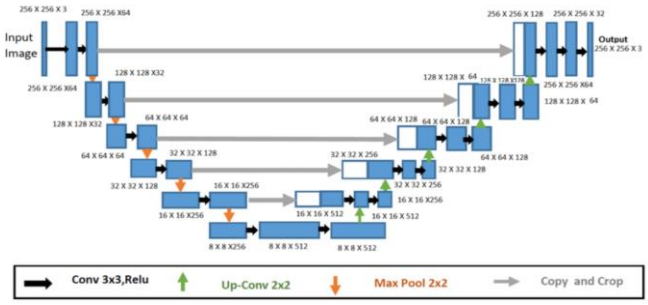


Fig. 7. Architecture of U-Net.

5) DeepLabV3+: The DeepLabV3+ model architecture was created especially for problems involving image segmentation. The encoder, which is represented by the blue line, and the decoder, which is represented by the red path, are its two primary parts. The input image is first processed using atrous (dilated) convolution in the encoder section in order to collect more spatial information without sacrificing resolution. A number of convolution and pooling operations are then performed, along with 1×1 convolutions to lower dimensionality and boost the effectiveness of feature representation. Low-level characteristics from previous network layers are combined with the encoder output in the

decoder stage after it has been upsampled by a factor of four and modified using 1×1 convolutions. The final segmentation map is created by upsampling the combined features by a factor of four after they have been improved by 3×3 convolutions (see Fig. 8).

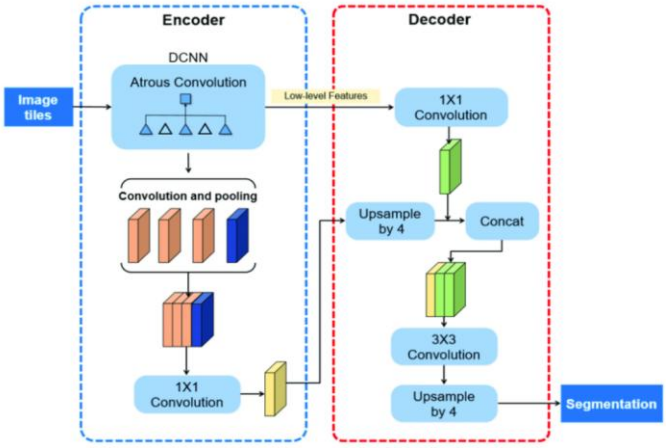


Fig. 8. Architecture of DeepLabV3+.

D. Lightweight Fusion Design for Segmentation

Additionally, a lightweight fusion approach was applied by combining the output probability maps of the two baseline models (U-Net and DeepLabV3+) using a simple adaptive weighting scheme. This approach aims to enhance segmentation consistency without significantly increasing architectural complexity or the number of parameters. The fusion was performed at the output probability level (mask probability map) by combining the probability maps from both models using a simple linear weighting scheme:

$$\text{Pfusion} = \alpha P_{U-net} + (1-\alpha)\text{P}_{\text{DeepLabV3+}}$$

where, α=0.5 serves as a balance weight between the two models. The fused result was then converted into the final binary mask using a threshold value of 0.5.

E. Training and Testing Scheme

The models were trained using 80% of the dataset for training and 20% for testing. The Adam optimizer [1] was employed with a learning rate of 0.001 and a batch size of 32 over 50 epochs. For classification, the categorical cross-entropy loss was utilized due to its effectiveness in multi-class settings [2]. For segmentation tasks, a composite loss combining Dice Loss and Binary Cross-Entropy (BCE) was applied to balance region overlap and pixel-wise prediction accuracy [3-4].

F. Performance Evaluation

The classification and segmentation of images of tropical fruits were the two main tasks for which performance evaluation was carried out. We used the confusion matrix, F1-score, recall, accuracy, and precision as evaluation metrics for the classification job. While precision and recall indicate the model's capacity to accurately identify positive occurrences and discover all pertinent instances, respectively, accuracy gauges the overall correctness of predictions. Particularly helpful when there is a class imbalance, the F1-score offers a harmonious compromise between recall and precision [33-34]. The performance of VGG16, ResNet50, and EfficientNet-B0 in classifying images of strawberries and pineapples was evaluated and compared using these criteria.

The evaluation used pixel accuracy, Dice Similarity Coefficient (DSC), and Intersection over Union (IoU) for the segmentation job. Because IoU and DSC can measure the spatial overlap between predicted regions and ground truth, they are frequently utilized in semantic segmentation tasks [35]. The percentage of correctly categorized pixels throughout the entire image is known as pixel accuracy. The segmentation performance of U-Net and DeepLabV3+ in recognizing fruit regions and detecting areas of visual deterioration or damage was assessed using these criteria [36].

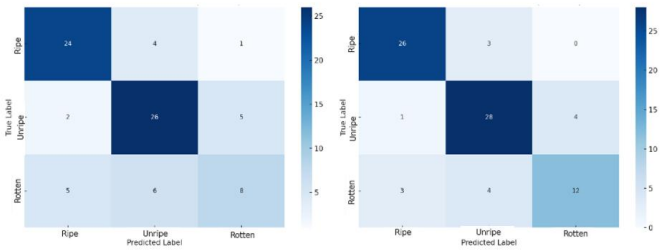
Besides classification accuracy, computational efficiency and statistical reliability are also crucial factors to ensure the feasibility of deploying the model in real-time or edge-based applications. The evaluation considers several complementary approaches, including inference time (runtime), number of parameters, and model size, to assess both predictive performance and computational efficiency of each architecture. Furthermore, a statistical significance test was conducted to verify that the performance differences among models are statistically meaningful.

To ensure model stability and reliability, a 5-fold cross-validation strategy was applied to both the pineapple and strawberry datasets. This approach allows for evaluating the model's generalization capability across different data variations while minimizing bias caused by uneven data partitioning.

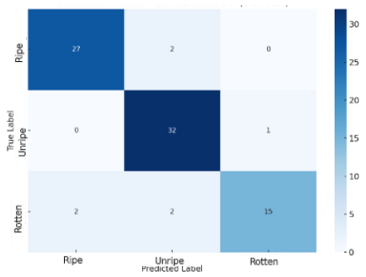
IV. Results and Discussion

A. Fruit Image Classification Results

In this horticultural fruit classification experiment (pineapples and strawberries), three CNN architectures were used: VGG16, ResNet50, and EfficientNet-B0. The dataset was divided into training, validation, and test data with a 70:15:15 ratio. To ensure model stability, 5-fold cross-validation was performed on both datasets. The experimental results, which illustrate the classification distribution and prediction error patterns, are shown in the Fig. 9 and Fig. 10 below. The confusion matrix was generated from the fold with the highest accuracy during the 5-fold cross-validation process.

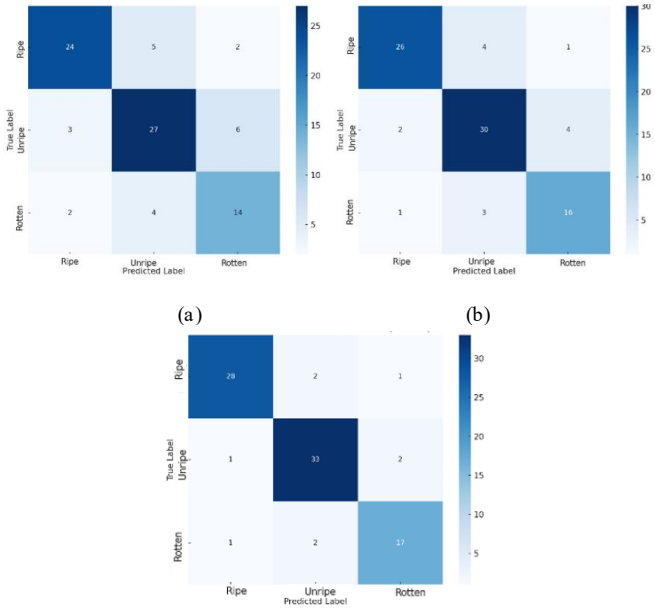


(a) (b)



(c)

Fig. 9. Confusion matrix: (a) VGG 16, (b) Resnet 50, (c) EfficientNet on pineapple dataset.



(a) (b)

(c)

Fig. 10. Confusion matrix: (a) VGG 16, (b) Resnet 50, (c) EfficientNet on strawberry dataset.

Each model was thoroughly evaluated on five different combinations of training and testing data, so that the results obtained were independent of a single data split. This evaluation included four main metrics: accuracy, precision, recall, and F1-score. The results of the average classification performance comparison of each architecture on the pineapple and strawberry datasets are shown in Table II and Table III. Furthermore, to illustrate the stability of inter-fold performance, the average value and standard deviation (mean ± SD) of each metric are also shown.

TABLE II. AVERAGE 5-FOLD CROSS-VALIDATION RESULTS ON PINEAPPLE DATASET

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VGG16	82.30 ± 1.1	80.40 ± 1.3	89.10 ± 1.5	80.50 ± 1.4
ResNet50	84.20 ± 0.9	85.00 ± 1.1	83.20 ± 1.0	84.10 ± 1.1
EfficientNet-B0	90.70 ± 0.7	91.10 ± 0.8	90.60 ± 0.6	91.20 ± 0.7

TABLE III. AVERAGE 5-FOLD CROSS-VALIDATION RESULTS ON STRAWBERRY DATASET

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VGG16	83.10 ± 1.0	82.50 ± 1.2	80.90± 1.1	82.00 ± 1.2
ResNet50	85.80 ± 0.8	85.90 ± 1.0	85.40 ± 0.9	85.70 ± 1.0
EfficientNet-B0	91.40 ± 0.6	91.80 ± 0.7	90.20 ± 0.6	90.60 ± 1.6

Table II and Table III summarize the model performance results by averaging across 5k folds. Based on these tables, it can be observed that EfficientNet-B0 consistently achieves the best performance in terms of accuracy, precision, recall, and F1 score on both datasets, followed by ResNet50 and VGG16. VGG16 has the lowest accuracy with a slightly higher standard deviation, indicating the model's sensitivity to variations in lighting and color on the fruit surface. Overall, the results on both datasets confirm that a more modern and lightweight architecture, such as EfficientNet-B0, is able to provide the best combination of high accuracy and inter-fold performance stability. The low standard deviation value of this model also indicates high reliability and good potential for application in fruit image classification systems in real environments.

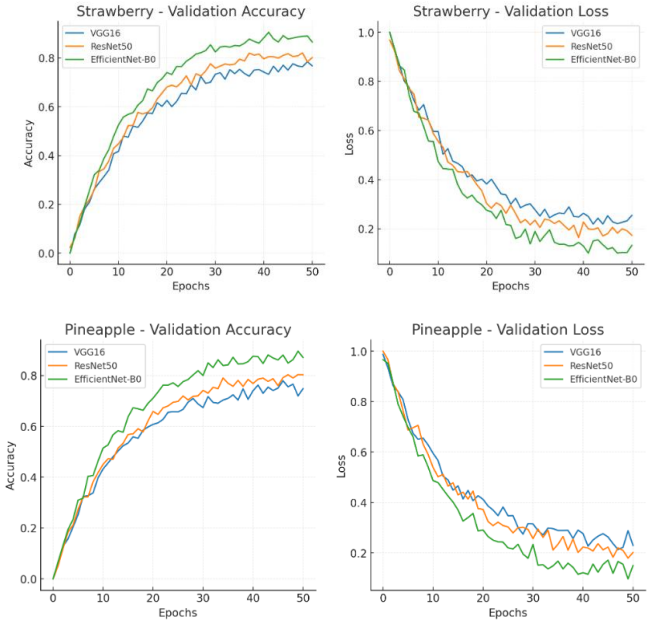


Fig. 11. Validation accuracy and validation loss: strawberry and pineapple.

To illustrate the performance of the models during the training process, the validation accuracy and loss curves on the strawberry and pineapple datasets are visualized. Fig. 11 presents the learning trends of each model (VGG16, ResNet50, and EfficientNet-B0) across 50 epochs.

The experimental results on the strawberry dataset show a stable trend of increasing validation accuracy as the number of epochs increases, reaching convergence around epoch 40. The EfficientNet-B0 model consistently achieved the highest accuracy, followed by ResNet50 and VGG16. This performance demonstrates that modern architectures such as EfficientNet are able to extract strawberry visual features more

effectively, particularly in distinguishing variations in color and texture that are crucial indicators for determining the condition of fresh and spoiled fruit. Meanwhile, on the pineapple dataset, all three models also exhibited consistent patterns of increasing validation accuracy and decreasing loss throughout the training process. EfficientNet-B0 once again achieved the best performance. These findings indicate that the more complex visual characteristics of pineapples, such as scaly skin patterns and lighting variations, are more effectively handled by the EfficientNet architecture. This is attributed to the optimization strategy employed by EfficientNet, which balances network depth, width, and resolution, thereby enabling richer and more accurate feature representations compared to ResNet50 and VGG16.

B. Segmentation Results

Segmentation experiments were conducted using U-Net and DeepLabV3+ on strawberry and pineapple images. The evaluation was carried out using three main metrics: Intersection over Union (IoU), Dice Coefficient, and mean Intersection over Union (mIoU). Table IV summarizes the performance of both models:

TABLE IV. EVALUATION METRICS AND AVERAGE SEGMENTATION PERFORMANCE (MEAN ± SD) ON STRAWBERRY AND PINEAPPLE DATASETS

Fruit	Model	IoU (%)	Dice (%)	mIoU (%)
Strawberry	U-Net	88.5 ± 0.9	89.5 ± 0.8	87.6 ± 1.0
Strawberry	DeepLabV3+	91.8 ± 0.7	91.5 ± 0.6	90.7 ± 0.8
Pineapple	U-Net	87.8 ± 1.0	88.7 ± 0.9	86.9 ± 1.1
Pineapple	DeepLabV3+	90.6 ± 0.8	91.7 ± 0.7	89.5 ± 0.9

The average segmentation performance results in Table IV show that the DeepLabV3+ model consistently produces higher IoU, Dice, and mIoU values than the U-Net for both datasets: strawberry and pineapple. The performance difference between the two models appears stable, as evidenced by the relatively small standard deviation values (ranging from ± 0.6 to ± 1.1), indicating good interfold consistency of the segmentation results in the 5-fold cross-validation scheme. DeepLabV3+'s superiority lies primarily in its use of Atrous Convolution and Atrous Spatial Pyramid Pooling (ASPP), which are highly effective in capturing variations in texture and object size on the fruit surface.

C. Computational Efficiency Aspect

In addition to accuracy and segmentation performance, computational efficiency is also an important consideration to ensure the feasibility of model deployment on systems with limited hardware resources. The efficiency evaluation is conducted by examining several key metrics, namely inference time (runtime), number of parameters, and model size. The results of the computational efficiency evaluation for all classification and segmentation models are presented in Table V.

TABLE V. COMPUTATIONAL EFFICIENCY EVALUATION OF CLASSIFICATION AND SEGMENTATION MODELS

Model	Task	Inference Time (s)	Number of Parameters (M)	Model Size (MB)
VGG16	Classification	0.042	138.3	528.0
ResNet50	Classification	0.038	25.6	98.0
EfficientNet-B0	Classification	0.026	5.3	20.4
U-Net	Segmentation	0.087	31.0	122.0
DeepLabV3+	Segmentation	0.093	43.5	175.0

D. Segmentation Results with Lightweight Fusion

This study applied a lightweight fusion approach to enhance segmentation consistency without increasing architectural complexity. The approach combines the output probability maps (mask probability maps) of two baseline models—U-Net and DeepLabV3+—through a simple linear weighting scheme. Table VI presents the segmentation performance after applying lightweight fusion between U-Net and DeepLabV3+:

TABLE VI. COMPARATIVE SEGMENTATION PERFORMANCE OF U-NET, DEEPLABV3+, AND FUSION MODEL ON STRAWBERRY AND PINEAPPLE IMAGES

Fruit	Model	IoU (%)	Dice (%)	mIoU (%)
Strawberry	U-Net	89.3	90.8	88.6
Strawberry	DeepLabV3+	92.1	93.5	91.7
Strawberry	Fusion (U+D)	**92.6**	**93.9**	**92.2**
Pineapple	U-Net	88.7	90.1	87.9
Pineapple	DeepLabV3+	91.4	92.7	90.8
Pineapple	Fusion (U+D)	**91.9**	**93.1**	**91.3**

The results demonstrate a consistent improvement across all key metrics after fusion. The average IoU increased by approximately 0.5–0.7%, while the Dice coefficient improved by about 0.4%. This enhancement is attributed to the complementary strengths of the two models: DeepLabV3+ excels in handling complex regions, whereas U-Net provides smoother boundary delineation. The combination produces more stable and accurate segmentation outcomes. This approach aligns with recent research trends emphasizing the effectiveness of lightweight hybrid models in improving segmentation robustness for agricultural imagery under natural illumination conditions [37].

E. Comparative Analysis

To ensure that the performance differences among models were not merely due to random variations in the data, a statistical significance analysis was conducted based on the 5-fold cross-validation results for both classification and segmentation tasks. This test aimed to verify the statistical validity of each model's performance, ensuring that higher-performing architectures exhibit improvements that are mathematically significant rather than coincidental.

The analysis was applied to both classification models (VGG16, ResNet50, and EfficientNet-B0) and segmentation

models (U-Net and DeepLabV3+) using significance levels of $\alpha = 0.05$ and 0.01. The complete results of the significance tests for both datasets are presented in Table VII and Table VIII below.

TABLE VII. STATISTICAL SIGNIFICANCE TEST RESULTS (5-FOLD CROSS-VALIDATION): CLASSIFICATION ON PINEAPPLE AND STRAWBERRY DATASETS

Model Comparison	Pineapple Dataset (p-value)	Strawberry Dataset (p-value)	Remark
VGG16 vs ResNet50	0.041 *	0.038 *	Significant
VGG16 vs EfficientNet-B0	0.007 **	0.005 **	Highly significant
ResNet50 vs EfficientNet-B0	0.018 *	0.015 *	Significant

TABLE VIII. STATISTICAL SIGNIFICANCE TEST RESULTS (5-FOLD CROSS-VALIDATION): SEGMENTATION ON PINEAPPLE AND STRAWBERRY DATASETS

Model Comparison	Pineapple Dataset (p-value)	Strawberry Dataset (p-value)	Remark
U-Net vs DeepLabV3+	0.009 **	0.006 **	Highly significant

The results of the statistical significance test confirm that the observed differences in performance between models are not due to random variations in the data, but rather have real statistical significance. The overall experimental results confirm that EfficientNet-B0 consistently outperformed ResNet50 and VGG16 in classification tasks, demonstrating an excellent balance between parameter efficiency and generalization capability. This finding aligns with the report, where EfficientNet achieved up to 97% accuracy in grape and potato leaf disease classification [38]. A follow-up study also emphasized the superior efficiency of EfficientNet compared to other architectures [39].

For segmentation tasks, DeepLabV3+ proved more precise than U-Net, particularly in handling the complex textures of strawberries and pineapples—the multiscale effect of atrous convolution and the ASPP module played a pivotal role. Evidence from the SugarBeets study also showed that DeepLabV3+ achieved a higher mIoU compared to U-Net [40].

Nevertheless, some error patterns persisted. False positives frequently occurred in decayed regions obscured by shadows, while false negatives often appeared when small decayed spots were hidden, similar to the findings on apples [41]. Segmentation also exhibited mask leakage along object boundaries, as reported in tomato images [42].

As a direction for further development, integrating adaptive multitask architectures such as AISAM-CSNet (Artificial Immune System-controlled Adaptive Multi-agent Classification and Segmentation Network) appears promising for improving robustness. AISAM-CSNet combines multi-agent CNNs, reinforcement learning, and optimization inspired by the Artificial Immune System, consistent with current trends in multitask learning and biologically inspired optimization within agricultural computer vision [43].

V. CONCLUSION

The experimental results demonstrated that deep learning methods are highly effective for assessing the visual quality of tropical fruits. Among the evaluated models, EfficientNet-B0 consistently achieved superior performance in the classification tasks, with an average accuracy of 90.7% for pineapples and 91.4% for strawberries, outperforming both ResNet50 and VGG16. This result confirms that lightweight and modern architectures are more capable of capturing discriminative visual patterns related to fruit ripeness and surface damage compared to conventional CNNs. Meanwhile, for the segmentation task, DeepLabV3+ outperformed U-Net, with an IoU of 90.6% for pineapple and 91.8% for strawberry, while the Dice coefficient was 91.5% for strawberry and 91.7% for pineapple. These results confirm that both architectures can serve as a reliable basis for automatically detecting ripeness and damaged areas in fruit.

Beyond numerical achievements, the findings highlight the practical contribution of deep learning in accelerating postharvest quality inspection while providing an empirical foundation for developing more adaptive multitask architectures. This has significant implications for the horticulture industry by reducing reliance on labor-intensive manual inspection.

For future research, the integration of classification and segmentation into a single multitask framework, such as AISAM-CSNet (Artificial Immune System-Controlled Adaptive Multi-agent Classification and Segmentation Network), holds strong potential to further enhance accuracy and adaptability under diverse real-world conditions. AISAM-CSNet combines classification and segmentation in a multi-agent reinforcement learning framework, with hyperparameter optimization inspired by the Artificial Immune System. Such an approach is expected to improve adaptability and accuracy in tropical fruit quality assessment systems, thereby supporting supply chain efficiency and ensuring consistent fruit quality.

REFERENCES

[1] Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. Computers and Electronics in Agriculture, 147, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

[2] Sudars, K., Pathak, K., Ferentinos, K. P., Tzounis, M., & Bochtis, A. (2020). Deep learning for image-based fruit classification and quality evaluation: A review. Agronomy, 10(6), 886. https://doi.org/10.3390/agronomy10060886

[3] Islam, M. A., Ali, M. S., & Hasan, M. M. (2023). Real-time fruit detection and classification using improved YOLOv5 and ResNet50. Computers and Electronics in Agriculture, 205, 107609. https://doi.org/10.1016/j.compag.2023.107609

[4] Campos, P., Marañón-Blanco, J. U., et al. (2023). Fruit quality classification using VGG16 and ResNet50 on FruitNet dataset. Computers and Electronics in Agriculture, 203, 107472. https://doi.org/10.1016/j.compag.2023.107472

[5] Li, X., Liu, Y., Zhang, H., & Wu, L. (2024). EfficientNet for plant leaf classification. Computers and Electronics in Agriculture, 209, 107926. https://doi.org/10.1016/j.compag.2024.107926

[6] Hasan, M. M., Tumpa, A., Mamun, A. M. A., & Shamsuddin, M. A. (2021). Fruit image classification using deep learning and Gaussian process regression. Information Processing in Agriculture, 8(3), 414–424. https://doi.org/10.1016/j.inpa.2020.08.004

[7] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1409.1556

[8] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (pp. 6105–6114). PMLR. https://proceedings.mlr.press/v97/tan19a.html

[9] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In MICCAI (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

[10] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In ECCV (pp. 833–851). Springer. https://doi.org/10.1007/978-3-030-01234-2_49

[11] Wang, Z., Hu, M., & Zhai, G. (2020). Apple segmentation based on DeepLabV3+ with dense conditional random field. Computers and Electronics in Agriculture, 179, 105842. https://doi.org/10.1016/j.compag.2020.105842

[12] Mo, C., Ma, C., & Wang, Y. (2024). Ripeness detection of sugar apple using MobileNetV2-DeepLabV3+. Agronomy, 14(4), 591. https://doi.org/10.3390/agronomy14040591

[13] Chiu, M. T., Tao, A., et al. (2020). Agriculture-Vision: A large aerial image database for agricultural pattern analysis. In CVPRW (pp. 2870–2879).https://openaccess.thecvf.com/content_CVPRW_2020/html/w41/Chiu_Agriculture-Vision_CVPRW_2020_paper.html

[14] Shorten, A., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6, 60. https://doi.org/10.1186/s40537-019-0197-0

[15] Chen, Y., Wang, S., Xu, X., & Xie, Y. (2021). Domain adaptation for plant disease recognition with adversarial training. Computers and Electronics in Agriculture, 187, 106303. https://doi.org/10.1016/j.compag.2021.106303

[16] Rahnemoonfar, M., & Sheppard, C. (2017). Deep count: Fruit counting based on deep simulated learning. Sensors, 17(4), 905. https://doi.org/10.3390/s17040905

[17] Arif, M., Fatima, S., & Khan, M. A. (2022). Real-time citrus fruit classification using deep convolutional neural networks. Sustainable Computing: Informatics and Systems, 35, 100753. https://doi.org/10.1016/j.suscom.2022.100753

[18] Wagle, K., Dinh, A., & Wang, Z. (2023). Edge deployment of EfficientNet for fruit classification. Computers and Electronics in Agriculture, 205, 107638. https://doi.org/10.1016/j.compag.2023.107638

[19] Reyes, R., Pérez, J., & Soto, A. (2022). Fine-grained fruit classification using EfficientNet and transfer learning. Journal of Imaging, 8(2), 50. https://doi.org/10.3390/jimaging8020050

[20] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4700–4708). https://doi.org/10.1109/CVPR.2017.243

[21] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. In CVPR (pp. 4510–4520). https://doi.org/10.1109/CVPR.2018.00474

[22] Rauf, H. T., Lali, M. I. U., & Khan, M. A. (2022). Performance comparison of CNN architectures for fruit classification. Artificial Intelligence in Agriculture, 6, 24–32. https://doi.org/10.1016/j.aiia.2022.05.001

[23] Fang, J., Wang, C., & Xu, T. (2022). Strawberry segmentation using attention U-Net in field conditions. Computers and Electronics in Agriculture, 198, 107061. https://doi.org/10.1016/j.compag.2022.107061

[24] Jamil, N., Rauf, H. T., & Mahmood, T. (2023). Mango fruit segmentation using squeeze-excitation U-Net. Journal of Food Measurement and Characterization, 17, 2023–2031. https://doi.org/10.1007/s11694-023-01474-6

[25] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2961–2969). https://doi.org/10.1109/ICCV.2017.322

[26] Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep High-Resolution Representation Learning for Visual Recognition. In CVPR (pp. 5390–5399). https://doi.org/10.1109/CVPR.2019.00553

[27] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(12), 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

[28] Milioto, A., Lottes, P., & Stachniss, C. (2018). Real-time semantic segmentation of crop and weed for precision agriculture robots leveraging background knowledge in CNNs. In IROS (pp. 2229–2235). https://doi.org/10.1109/IROS.2018.8593940

[29] Singh, D., Jain, N., & Nandi, G. C. (2020). Fruit recognition in natural environment using deep learning. Procedia Computer Science, 167, 2201–2210. https://doi.org/10.1016/j.procs.2020.03.270

[30] Xu, L., Ma, R., Zhang, C., & Hu, W. (2021). Challenges in cross-domain apple classification. Sensors, 21(2), 509. https://doi.org/10.3390/s21020509Rahnemoonfar, M., & Sheppard, C. (2017). Deep count: Fruit counting based on deep simulated learning. Sensors, 17(4), 905. https://doi.org/10.3390/s17040905

[31] Adhil, P. K. (2022). *Pineapple* [Data set]. Kaggle. https://www.kaggle.com/datasets/adhilpk/pineapple

[32] Basit, A. (2023). *Strawberry dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/abdulbasit31/strawberry-dataset

[33] Zhang, Y., et al. (2019). Improved cross-entropy loss function for deep neural networks classification. IEEE Access, 7, 64366–64374. https://doi.org/10.1109/ACCESS.2019.2917624

[34] Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-Net: Fully convolutional neural networks for volumetric medical image segmentation. Proceedings of the 4th International Conference on 3D Vision (3DV), 565–571. https://doi.org/10.1109/3DV.2016.79

[35] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Cardoso, M. J. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In Deep learning in medical image analysis and multimodal learning for clinical decision support (pp. 240–248). Springer. https://doi.org/10.1007/978-3-319-67558-9_28

[36] Akther, J., Harun-Or-Roshid, M., & Islam, A. (2024). Comparative analysis of CNN, EfficientNet and ResNet for grape and potato leaves disease prediction: A deep learning approach. ResearchGate. https://10.21203/rs.3.rs-5037532/v1

[37] Zhang, L., Li, M., Zhang, P., & Liu, P. (2025). EfficientSegNet: Lightweight Semantic Segmentation with Multi-Scale Feature Fusion and Boundary Enhancement. Sensors, 25(19), 5934.

[38] Kansal, K. (2024). ResNet-50 vs. EfficientNet-B0: Multi-Centric Classification Performance Comparison. Procedia Computer Science. https://doi.org/10.1016/j.procs.2024.04.007

[39] Liu, C. (2025). DeepLabV3+ performance on SugarBeets dataset. Technical Report, DIVA Portal. https://liu.diva-portal.org/smash/get/diva2%3A1973627/FULLTEXT01.pdf

[40] Rathore, D., Divyanth, L. G., Reddy, K. L. S., Chawla, Y., Buragohain, M., Soni, P., ... & Ghosh, A. (2023). A two-stage deep-learning model for detection and occlusion-based classification of Kashmiri orchard apples for robotic harvesting. Journal of Biosystems Engineering, 48(2), 242-256. . https://link.springer.com/article/10.1007/s42853-023-00190-0

[41] Shoaib, M., Hussain, T., Shah, B., Ullah, I., Shah, S. M., Ali, F., & Park, S. H. (2022). Deep learning-based segmentation and classification of leaf images for detection of tomato plant disease. Frontiers in plant science, 13, 1031748. https://doi.org/10.3389/fpls.2022.1031748

[42] Wang, Y., Yang, L., Liu, X., & Yan, P. (2024). An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3+https://www.nature.com/articles/s41598-024-60375-1

[43] Moisés, A. G., et al. (2023). Augmentation techniques for deep learning in agrifood applications. Sensors, 23(20), 8562. https://doi.org/10.3390/s23208562

Towards Designing a Blockchain-Based Model for E-Book Publishing

Maznun Arifa Mohammadan Makhtar[1], Novia Admodisastro[2], Suleymenova Laura Askarbekkyzy[3]
Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Selangor, Malaysia[1,2]
South Kazakhstan Pedagogical University, Shymkent, Kazakhstan[3]

***Abstract*—This paper examines the application of blockchain technology in e-book publishing by analyzing previous research and identifying current limitations. The study investigates how smart contracts and cryptographic algorithms can facilitate agreements between publishers and authors. While blockchain has been widely adopted in digital publishing domains, such as image, video, music, and scientific journals, research on its application to e-books remains limited. Existing solutions typically address individual challenges such as transaction transparency, authenticity, or copyright protection, but rarely integrate them into a single framework. To provide a systematic synthesis of prior works, this paper develops a taxonomy of blockchain-based e-book publishing models across six dimensions: platform, storage, smart contract usage, cryptographic algorithm, tokenization, and actors. This paper reviews seven (7) blockchain-based models in e-book publishing and identifies their limitations. Based on these insights, a conceptual blockchain-based smart contract model for e-book publishing was proposed using Ethereum platform, incorporating InterPlanetary File System (IPFS) storage and cryptographic algorithms. The proposed model has the potential to significantly enhance the security and rights protection for authors and publishers, thereby fostering a more secure and equitable e-book publishing landscape.**

Keywords*—*e-book publishing; Ethereum; blockchain technology; smart contract

I. Introduction

The book publishing industry has been a central component in the creation and dissemination of knowledge through books, magazines, journals, articles, dictionaries, and other reading materials. As technology evolves, many people read and distribute digital content online. However, digital publishing is vulnerable to illegal downloads, copying, content sharing, and piracy. Some issues regarding copyright law in the digital environment are a lack of transparency about the legal status of copyrighted works, piracy, and difficulties in getting compensated fairly [1]. This scenario also occurs in the publishing industry, where publishers need to ensure the e-book is secure while protecting the author's and publisher's rights.

An electronic book, also known as an e-book, is a digital version of traditional reading materials. The term "e-book" has been changing depending on the way it is presented and the evolution of technology [2]. Recent research in e-book publishing shows that non-transparent transactions, unfair royalty distribution, unprotected copyright, and unauthenticity of published materials are the main problems for digital content creators in this era. E-book publishing has created a trustless environment, as nobody is responsible for the security, validity, and ownership of the purchased e-book [3]. The authors do not have complete control over their e-books after publication. Non-transparent transactions may deny the actual sales reports to the authors [4]. The e-book's sales history is difficult to track and not tamper-proof, leaving it vulnerable to illegitimate alteration and unauthorized distribution. Unprotected copyright protection may lead to unfair royalty payments because the authors were unable to track each sales transaction [5]. In addition, published e-books lack authenticity because their content can be replicated, reproduced, and redistributed unethically. Unauthenticated e-books may affect total sales and the royalty paid to the authors [6].

Blockchain has significantly contributed to digital publishing, including image, video, music, scientific journals, and e-book publishing. The concept of blockchain was initially proposed by Satoshi Nakamoto in 2008. Nakamoto introduced Bitcoin, a cryptocurrency that allows a peer-to-peer electronic cash system [7]. Blockchain is a data structure that operates in a decentralized, peer-to-peer network consisting of a chain of blocks. Each block contains transactions and also a hash value from the previous block, which makes it immutable and tamper-proof [8]. Distributed ledger technology (DLT) is the main characteristic that enables all transactions to be stored in a blockchain [9]. DLT allows data exchange between the network participants without intermediaries [10]. Blockchain works on a consensus algorithm, a mechanism used among the participants (nodes) to agree on the transactions [11]. Blockchain creates an audit trail of transactions between participants. Therefore, payment can be made directly without the intervention of a third party, such as a financial institution, because blockchain can act as a mediator for data transformation or calculation [12]. When blockchain is combined with smart contracts and cryptography, it can help solve problems of trust, authenticity, and copyright protection in e-book publishing [6].

This paper examines prior work to gain a deeper understanding of the design, methods, algorithms, limitations, and weaknesses. A blockchain-based conceptual model for e-book publishing using the Ethereum platform was proposed to address all those challenges. The model incorporates the InterPlanetary File System (IPFS) and a smart contract to provide three key benefits: transparent transactions, e-book authenticity, and copyright protection to the authors through fair royalty distribution.

Unlike the previous study, which focuses on a single or two aspects such as transparency of transactions, authenticity, or copyright protection, this study integrates all three elements into a single conceptual model. The combination of IPFS,

cryptographic mechanisms, Ethereum smart contract, and multi-actor involvement reflects the structure of the real e-book publishing ecosystem. This holistic approach constitutes the core novelty of this work and directly addresses the identified research gap, thus adding value to the current body of knowledge.

The rest of this paper is structured as follows: Section II presents the research contribution; Section III presents the background information; Section IV reviews related works; Section V introduces the taxonomy of blockchain-based e-book publishing, and Section VI explains the proposed conceptual model; Section VII presents the discussion, and finally, Section VIII concludes with future work.

II. Research Contribution

This study contributes to the domain of blockchain-based e-book publishing by addressing the specific challenges of e-book distribution. While prior work has proposed solutions focusing on isolated aspects, such as transparency, authenticity, or copyright protection, these solutions rarely integrate these elements into a unified framework. The contributions of this paper are:

- Conceptual model for e-book publishing: The proposed model integrates Ethereum smart contracts with InterPlanetary File System (IPFS) storage and cryptographic algorithms to ensure transparent transactions, authenticity verification, and copyright protection.
- Multi-actor involvement: unlike previous work that primarily emphasizes author-customer interactivity or self-publishing, this model incorporates the publisher and the author as key participants alongside the customer. This reflects the practical realities in the e-book publishing industry.
- Comparative analysis and gap identification: Through a comparative review of seven (7) related blockchain-based models, this study identifies key limitations in existing approaches. The proposed model explicitly addresses these gaps by combining transaction transparency, e-book authenticity, and automated royalty payment within a single conceptual framework.
- Development of a taxonomy: This paper develops a taxonomy that classifies blockchain-based e-book publishing models across six dimensions: platform, storage, smart contract usage, cryptographic algorithms, tokenization, and actors. The taxonomy provides a structured framework for comparing prior works and identifying research gaps, thereby establishing a theoretical reference for future studies.

III. Background

A. Electronic Book (E-Book) Publishing

The first e-book development started in the 1970s by Project Gutenberg and the Oxford Text Archive. Project Gutenberg was initiated by Michael Hart in 1971 at the University of Illinois, and the first e-book published was "Declaration of Independence," followed by "Bill of Rights" in 1971 [13]. These publications led to Hart's other big idea: to create the world's first e-book library. In 1976, the Oxford Text Archive focused on providing electronic academic resources to scholars. From the 1980s to the 1990s, the development of e-books expanded to several formats and media. The publishers gain more confidence in publishing e-books for academic areas and linking with libraries as a niche market. In 1989, e-books were published on Compact Disc Read-Only Memory (CD-ROM) and could be accessed using computers [14]. Since then, several discussions about digital copyright have started.

E-book publishing experienced massive growth alongside technological development. E-books can be accessed on multiple devices, such as PDAs, Pocket PCs, e-book readers, and mobile phones [15]. A new horizon of e-books started when Amazon introduced the Kindle in 2007. Kindle changes the reader's reading behavior because it provides readability flexibility [16]. Other e-book readers, such as Barnes & Noble Nook and Kobo, also entered the market due to their convenience and cost-effectiveness.

The potential and benefits of adopting blockchain technology into publishing, specifically intellectual property and monetary rights, have been addressed. The most important highlight is self-publishing 3.0, which means a direct transaction between the authors and readers without intervention from any third party [17]. Instead of self-publishing, some authors may need to publish their books through a publisher. In the e-book publishing process, the process can be classified into three (3) elements: content providers, service platforms, and reading devices [16]. The end-user is the customer who accesses the e-book by downloading it onto their reading devices. Generally, e-book publishing is the same as physical books, but with additional parties involved in digitization. The entire process of e-book publishing is illustrated in Fig. 1.

Fig. 1. E-book publishing process.

The parties involved in e-book publishing are as follows: 1) author: the person who writes a manuscript to be published by the publisher and owns the copyright of their respective work; 2) publisher: the organization or company that processes the author's manuscript to be published in digital format, manage copyright, receive sales payment and pay royalty to the author and 3) customer(s)/subscriber(s): the potential buyer of the e-book.

The publishing process begins when the author(s) write and send a manuscript to the publisher. The manuscript will be evaluated before approval for publication. The publisher will decide on the e-book format with the consent of the author(s). Then, the process of developing an e-book begins with editing and proofreading. Some of the e-books might need illustration and design. The programming part will be done to integrate the

multimedia elements and the system used. The final e-book product format varies, typically PDF or EPUB. The e-book will be published on various websites and mobile platforms, including the App Store, Google Play, and the Huawei App. The customer(s) or subscriber(s) will purchase the e-book from the selling platform and access it on their devices.

Previous related works [4], [5] have suggested that blockchain is a better solution for e-book publishing due to its characteristics, which can provide transparency, integrity, and immutability. The following section will discuss the explanations of blockchain, decentralized repository systems, and smart contract.

B. Blockchain

In recent years, blockchain technology has expanded far beyond cryptocurrencies. Blockchain has been applied in various industries, including healthcare, automotive, insurance, agriculture, advertising, and education [18]. Blockchain's decentralized peer-to-peer networks and distributed ledger technology features benefit businesses by providing transparency, immutability, security, and tamper-resistance. Blockchain will mitigate all risks of a "principal-agent" problem, which then shapes the blockchain character as "trustless" or "trust-free" [19].

Blockchain is a chain of blocks linked together using cryptographic algorithms that permanently record transactions in a large, distributed ledger. Each block stores the hash of the previous block, and a new block is created when the sender submits a transaction. This security feature in blockchain makes it extremely difficult to manipulate the data. The latest block will undergo mining before being added to the chain. Mining is a data validation process that involves solving a specific algorithm. Miners will validate and verify the transactions using various consensus protocol mechanisms, such as Proof of Work (PoW) or Proof of Stake (PoS). This protocol ensures that peers have the same data in the blockchain [8]. Every transaction in the blockchain is timestamped, and once added, it cannot be changed.

Depending on its objectives, each organization may choose the type of blockchain that best meets its requirements. Each type of blockchain has different characteristics and functionalities. There are three (3) main types of blockchain: public (permissionless) blockchain, private (permissioned) blockchain, and consortium/federated blockchain [20].

- A public blockchain is open-access, decentralized, and allows any node to enter the network. The nodes read, write, and validate the data through a consensus mechanism. No individual or organization owns the public blockchain, which is self-governing. Examples of public blockchains are Bitcoin, Ethereum, and Litecoin.
- A private blockchain is a centralized blockchain managed by an organization or authority. It is usually applied to small businesses and has limited use within private entities. Only authorized users can view the data and transactions within the network. Corda and Ripple are examples of private blockchains.
- Consortium or Federated blockchain is a combination of public and private blockchain in which a part of the blockchain is partially centralized. Only selected nodes can choose the services they intend to participate in, and the remaining nodes can only access transactions, not the consensus process. Hyperledger Fabric is an example of a Consortium blockchain.

Shukla [21] provided an overview of blockchain research and outlined a future agenda, highlighting how blockchain applications are expanding beyond cryptocurrency into domains such as copyright protection and publishing. This perspective reinforces the timeliness of applying blockchain in e-book publishing.

The following section will discuss the Ethereum blockchain and its components: the repository system, smart contracts, and cryptographic algorithms.

1) Ethereum blockchain: The Ethereum Blockchain is an open-source, public (permissionless) blockchain platform with smart contract functionality that can verify a predetermined set of rules. Any node can join the Ethereum network if the user has an Ethereum client account [12]. Fig. 2 illustrates the deployment of the e-book publishing model in the Ethereum blockchain by executing a smart contract on the Ethereum Virtual Machine. This decentralized machine can execute smart contract code in Ethereum nodes. A dApp is a decentralized application that acts as a web interface to connect users with a smart contract.

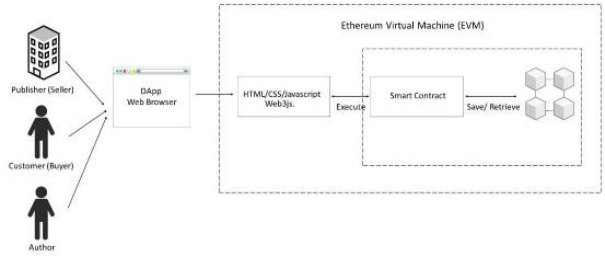


Fig. 2. Decentralized blockchain-based e-book publishing architecture (Nurgaida Yutia et al., 2022) adapted with permission.

In e-book publishing, a public (permissionless) blockchain is better suited for adaptation because the system requires involvement from both internal and external parties, such as publishers, authors, and customers. Furthermore, the system aims to facilitate a safe, secure trading environment for all stakeholders. Public blockchains provide high transparency, enabling authors and publishers to access sales and royalty records. Ethereum is an example of a public blockchain that offers functionality beyond cryptocurrency and has been widely used to develop decentralized applications (dApps) to support business management processes.

2) Smart contract: A smart contract is one of the main components of the Ethereum blockchain. A smart contract refers to a programmable agreement embedded in the blockchain that automatically enforces predefined rules between stakeholders. It works as a contract between two (2) or more parties that has been electronically transformed. A smart

contract allows code to specify how the process should be managed, and it provides a distributed, unchangeable record of all events that have occurred [18]. The key elements of smart contracts are the elimination of trusted third parties, forger assistance, transparency, autonomous execution, accuracy, and paperless motivation [14].

In e-book publishing, a smart contract is an agreement between the author(s) and the publisher. A smart contract can automate royalty payments without intermediaries. Publishing rights, such as the author's copyright and the publisher's ownership, can be securely recorded on the blockchain, making them verifiable and legitimate. Smart contract programmed using Solidity Programming Language on a web-based Integrated Development Environment (IDE), Remix. Smart contracts do not involve any third party to create an agreement between the author(s) and the publisher.

3) Repository system: Blockchain architecture is designed around a consensus mechanism, and decentralization requires full participation by nodes. The expansion in transaction volume will increase data storage in parallel with the involvement of nodes to maintain its stability [22]. Therefore, there is a need for an alternative way to store large amounts of data in decentralized storage systems such as InterPlanetary File System (IPFS), PFS, Storj, SWARM, Filecoin, and Sia.

IPFS is a globally distributed system widely used in the Ethereum Blockchain environment. While blockchain can provide high security, storing data on it would be very expensive. Currently, the Ethereum chain size is around 500GB – 1TB, and every node should be able to store the data. If the block expands, it would not be feasible for the blockchain to perform the transactions [23]. IPFS can store large files, and the data links or hashes are stored on the blockchain. A cryptographic algorithm in IPFS made it immutable. IPFS uses a content-addressing technique to store files. The distributed storage system architecture in IPFS guarantees the accessibility of the data on the network because all the peers have exact copies of the data [24]. If the data has been modified, IPFS will compute a new hash and update it across peers.

For example, Pubchain, a solution for academic resource publications in the Ethereum blockchain, uses IPFS for publication storage, was proposed by [25]. Authors will upload their articles to the system, and the articles will be time-stamped. The articles will be registered in the blockchain as permanent records and stored in IPFS, with metadata created accordingly. The smart contract will write the metadata into the blockchain.

4) Cryptographic algorithm: Cryptographic algorithms are the backbone of blockchain. Blockchain and smart contracts use various cryptographic algorithms to ensure the authenticity, integrity, and privacy of data. These algorithms are vital in securing transactions, ensuring immutability, and enabling blockchain functionality. Some algorithms that are commonly used are:

a) Hash functions: Cryptographic hash functions formed from the Merkle-Damgard method need one input: the message to be hashed [26]. A hash function is a one-way function that is not reversible. One input produces a single, fixed-length output, called a hash value. A hash function can ensure the integrity and authenticity of a message. Message Digest (MD) and Secure Hash Algorithms (SHA) are examples of hash functions.

b) Public-key cryptography: Encryption and decryption of a message using a pair of keys assigned to the corresponding receiver. The readable message will be encrypted with the receiver's public key and decrypted with the receiver's private key. Only the receiver can encrypt the message, thus providing confidentiality and authenticity. Public-key cryptography is the technology behind digital wallets in the blockchain [27]. Examples of public-key cryptography are Rivest-Shamir-Adleman (RSA) and Elliptic Curve Cryptography (ECC).

c) Digital signatures: Mathematical techniques for verification and authentication of a digital message by assigning a pair of keys, private-public keys, with the involvement of a signer and verifier [28]. A digital document will be signed using the sender's private key, and the verifier will verify the originator by decrypting it with the sender's public key. Digital signatures are often used with hash functions to verify the authenticity and integrity of a message.

C. Blockchain in Digital Publishing

Studies from 2020 to 2023 show that the work of blockchain models for digital publishing has mainly been developed for other assets, such as image publishing [29], music streaming [30], scientific publications [31], video publishing [32] and the student's academic record [33]. Nevertheless, the implementation of blockchain in e-book publishing has not been extensively explored. The existing work is unsuitable for e-books because its process flow and architectural design differ from those of e-books.

Therefore, this study will focus on implementing blockchain technology in e-book publishing to provide transparency, authenticity, and copyright protection. Blockchain can enable secure transactions and potentially create a new business model. Trusted distributed ledger technology is the primary reason blockchain should be used in the publishing industry. Every e-book published has its respective copyright owner. The publisher must retain the transaction records and reasonably manage the author's royalties.

In the blockchain e-book publishing scenario, there are five (5) practical use cases.

- Track of history records: Blockchain can record every transaction. The records are transparent and can be viewed by everyone; therefore, the origins of the transactions are traceable.
- Royalty payment systems: The Ethereum blockchain can give authors a fair royalty distribution by constructing a smart contract with pre-determined rules. The author must verify the smart contract before the e-book is up for sale. Royalty will be delivered automatically without the intervention of a third party, such as a financial institution or publisher. The author will receive the royalty after every successful customer purchase.

- Piracy prevention: e-Book publishing is easily exposed to various kinds of piracy, such as copying, illegal downloading, reprinting, and redistribution. Blockchain, using cryptographic algorithms, can mitigate the risks of any fraud.
- Copyright protection: A smart contract in blockchain enables fixed copyright registration and allows creative rights.
- Cost-effective for large data storage: e-Book size varies depending on the content. Some publishers use a large, centralized storage system, which is both costly and complex. Others might use the cloud, but costs will increase as data grows. However, blockchain and peer-to-peer storage systems, such as IPFS, can provide decentralized, secure storage at minimal cost.

IV. Related Works

This section describes seven (7) related works on e-book publishing using the blockchain model found in the literature. These works serve different objectives, involve various entities, utilize different platforms, employ distinct processes, and employ diverse methods.

IPFS-Blockchain-Based Authenticity of Online Publications proposed by [6] aim to provide authenticity and integrity of published digital content in different versions of original e-books (languages). The proposed model is demonstrated in a system that uses the Ethereum blockchain and smart contracts to ensure the authenticity and integrity of the e-book. The system utilized IPFS to store the data, as it is more cost-effective than blockchain storage. E-book hash data will be generated in IPFS and stored in the blockchain. The hash will be retrieved when the transaction matches it with the database. If the hash has been altered, the original content has been modified. By using this method, only authentic, original e-books will be sold to customers.

Another blockchain model proposed by study [5] protects the author's copyright and transparency in transactions by providing a transparent sales ledger that could be accessed by the publisher, author, and customers. The model shown in Fig. 3 was built on the Ethereum blockchain, using a smart contract to track e-book sales. Every transaction will be verified by consensus and digitally signed using Elliptic Curve Cryptography (ECC) before being stored in the blockchain. Both the author and the publisher must deposit twice the actual book price for double verification. After the authorization process, half of the deposit will be refunded to the authors, and the publisher will get a full refund and the sales percentage. The royalty payment will be made immediately to the author upon completion of the transaction, thereby eliminating the need for third-party involvement and ensuring high reliability, integrity, and security.

An intellectual property trading system, named JS Digital Asset Trading System, was developed for e-book trading [34]. Only authorized persons can access the e-book in this system; only authentic books can be sold to customers. The system used Asymmetric Key Encryption (AKE), in which the author uses his public key to encrypt the e-book before uploading it into the system. Then, when the customer purchases the e-book, the author will decrypt it with his private key. The e-book will be encrypted again using the customer's public key and sent to the customer. The customer needs to decrypt the e-book using their private key. This approach ensures the security and authenticity of the trading process. However, the study did not discuss any solutions for royalty distribution among publishers and authors.

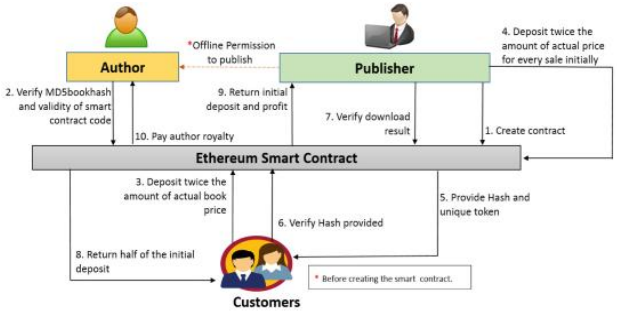


Fig. 3. Blockchain-based protection for author royalty of digital asset system architecture.

Another blockchain-based trading system scheme has been proposed by [35]. The JSP Digital Asset Trading System was developed to provide transparency, as the digital asset owner can view the transactions and receive payment immediately after the transaction (refer to Fig. 4). The "Random-Checker Proof of Stake" consensus mechanism was used to reduce transaction time, while AKE is used for verification. The proposed system uses a smart contract to allocate the share percentage between the authors. However, the proposed system does not involve the publisher.

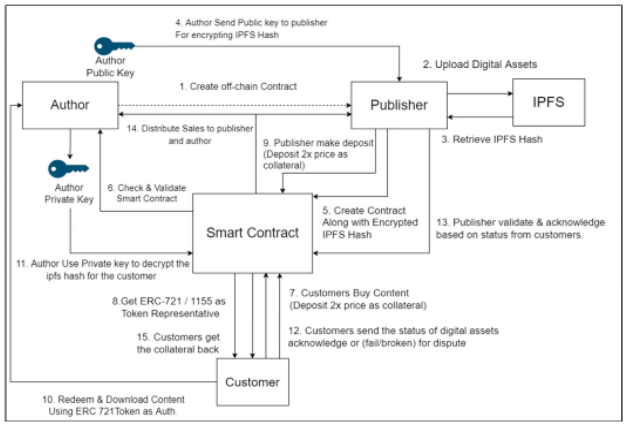


Fig. 4. Blockchain-based protection of author royalty of digital asset system overview.

A blockchain-based eBook market system proposed by [3] could facilitate self-publishing by enabling transparent transactions. The authors will receive direct payments from the system without intermediaries. In this system, a book can be sold by the chapter or as a whole. The proposed system uses a smart contract to ensure reliability, copyright protection, and the security of e-book transactions.

A Non-Fungible Token (NFT) is a unique identifier for digital assets that allows users to purchase ownership rights to those assets. NFTs are typically applied to collectible items, such as games, music, and art. A blockchain-based NFT model

was proposed by [36] for e-book publishing, it allows the author to sell individual chapters and the entire e-book. This model utilizes IPFS as an external storage system to reduce costs and energy consumption. An NFT e-book is linked with a smart contract to ensure security and preserve the owner's copyright. However, NFT item procurement uses a bidding system, meaning the potential buyer must bid on the e-book's price within a limited timeframe. The successful buyer will have ownership of the e-book if the author accepts the offer. Furthermore, NFT items are typically sold in limited quantities to preserve their uniqueness. Therefore, the publisher must create a unique e-book, for example, by offering an e-book with additional audio of the authors reading several chapters.

The last related work is by study [4] which enhances the earlier work of [5]. They proposed adding security features to the data links of smart contract digital assets using Elliptic Curve Cryptography (ECC). In earlier work, the data link for digital assets in smart contracts is in plain text, making it vulnerable to attack. Therefore, ECC is used to encrypt data links, thereby strengthening data integrity and authenticity. Smart contracts are also used for sales, payments, and royalty distribution between author and publisher. In addition, IPFS is used for data storage, and ERC-721 is used as a standard for token-based authorization. ERC-721 is a standard for non-fungible tokens on Ethereum, meaning each asset is unique.

A. Discussion of Related Works

Table I summarizes the comparison of the existing related work in blockchain-based e-book publishing. In conclusion, although they are in the same domain, each related work has its own primary objectives and is developed based on its specific needs and requirements. However, every work has its limitations and boundaries. Both [4], [5] proposed the same three (3) objectives: to provide transparency, to preserve copyright, and to ensure e-book authenticity. Although [6] and [34] proposed solutions for authenticity, their solution differs in their system architectures. The work by [6] involved authors, the main publisher, secondary publishers, and customers, whereas the work by [34] involved only authors and customers. Other work with the involvement of two (2) parties is by [35] and [3]. Both works suggested a service application for the author to self-publish their e-books. No publisher is involved in this process; thus, the purchasing transaction will be made only between the authors and customers. Another solution by [36] suggested non-fungible tokens (NFTs), a token representing a unique digital asset that could allow an individual to have ownership of the author's work. This model also involves the author and the customers. The variation in actor involvement highlights a critical distinction across existing models. Some approaches reflect a broader publishing ecosystem, while the solution involves direct interaction between customers and authors. Models that exclude publishers may be practical for independent distribution. Still, they do not capture the realities of e-book publishing, where publishers play a key role in validation, e-book management, and royalty distribution. The differences observed in actor participation further justify the need for structured classification. Therefore, to synthesize these variations and to better position the proposed conceptual model, this study introduces a taxonomy of blockchain-based e-book publishing models. The detail of the taxonomy is presented in Section V.

While the analysis in Table I highlights the comparison of the objectives, actor involvement, and limitations across existing works, revealing different publishing workflows and priorities, Table II complement by focusing on the technical implementation of these models, specifically the choice of blockchain platform, storage mechanisms, and cryptographic algorithms. Together, these tables offer a complete picture of both functional and technical perspectives of existing solutions.

Referring to Table II, four (4) works were built on the Ethereum blockchain platform [4], [5], [6], [34]. Publishers need a public environment and visibility to market e-books to potential buyers. Hence, Ethereum is the most suitable choice for publishers to trade their products. In [3], [36], the authors did not specifically mention which platform was used in their work, as the proposed solution generally works across any blockchain platform. [35] developed their blockchain platform and utilized their own cryptocurrency, JSP Coin.

However, six (6) out of seven (7) models use smart contract despite three (3) works not being built in Ethereum [3], [35], [36]. This is because the smart contract can also be deployed to other blockchain platforms. For example, a smart contract can be applied to Hyperledger Fabric using a "chaincode" package [37].

Regarding storage systems, four (4) works utilize IPFS, and three (3) works use other storage, such as local storage and blockchain storage. E-book files could be large, especially for books with illustrations and graphics. Storing e-books on the blockchain would be very costly; therefore, a better option would be a decentralized storage system, such as IPFS, Storj, or SWARM. Work by [5] applied cloud-based storage, but it could also be extended into other decentralized storage systems.

Related work shows that various cryptographic algorithms have strengthened the system's security. A hash function is the standard algorithm used because the hash is part of an element in the blockchain. If the e-book files are stored in IPFS, the IPFS hashes will be stored on the blockchain. The hash function is crucial to ensure the e-book is not modified during the transaction. Apart from that, [5] specifically mentioned the use of MD5Hash and IPFS hash in their solutions.

Digital signatures are also an essential feature of blockchain. A digital signature ensures the authenticity of the data transmission on the blockchain network. Various types of digital signatures were used in the study, including Elliptic Curve Cryptography (ECC), Elliptic Curve Digital Signature Algorithm (ECDSA), and Asymmetric Key Encryption (AKE). From the study, every solution used at least one digital signature method. The work by [3] is more comprehensive, which explains why a greater number of methods were employed in their solutions. A study by [36] explores the possibility of adapting non-fungible tokens (NFTs) to e-book publishing; specifically, the ERC-721 standard enables authors to create crypto assets. In e-book publishing, digital signatures are crucial to ensure that a real publisher uploads the authentic e-book. The customer will verify the e-book signature by decrypting it using the publisher's public key.

TABLE I. COMPARISON OF PREVIOUS RELATED WORKS OF BLOCKCHAIN IN E-BOOK PUBLISHING

No.	Authors	Proposed Work	Objectives	Actors	Limitations
1.	[6]	IPFS-Blockchain-Based Authenticity of Online Publications	Authenticity, Integrity	Author, publisher, customer	• Author creates smart contract, publisher validates it • No discussion on payment/royalty
2.	[5]	Blockchain-based Framework for Protecting Author Royalty of Digital Assets	Transparency, Copyright, Authenticity	Author, publisher, customer	• Both publisher and buyer must deposit 2× book price • May affect usability
3.	[34]	JS Digital Asset Trading System	Transparency, Authenticity	Author, customer	• 4× encrypt–decrypt process • 3× e-book transfers over network • Author must always be active to decrypt
4.	[35]	JSP Digital Asset Trading System	Transparency, Copyright	Author, customer	• Only author & buyer involved • Developed own blockchain platform • Uses its own cryptocurrency (JSP Coin) • Marketplace style, no publisher role
5.	[3]	Blockchain-based eBook Market System for Self-Publishing	Transparency, Copyright, Authenticity, Confidentiality, Validity, Integrity	Author, customer	• Designed for self-publishing only • Relies on third-party service application
6.	[36]	Non-Fungible Tokens (NFTs) for Publishing	Copyright	Author, customer	• NFT better for special/collectible editions • Buyer must bid in limited timeframe • Ownership can be resold
7.	[4]	Blockchain-based Protection of Author Royalty of Digital Assets	Transparency, Copyright, Authenticity	Author, customer	• Publisher & buyer must deposit 2× book price • Similar limitations to Nizamuddin et al. (2019)

TABLE II. SUMMARY OF STORAGE, PLATFORM, AND ALGORITHM USED IN BLOCKCHAIN-BASED E-BOOK PUBLISHING

No.	Authors	Smart Contract	Storage	Platform	Cryptographic Algorithm				Token / NFT
					Symmetric Encryption	Asymmetric Encryption	Digital Signature	Hashing / Integrity	
1.	[5]	✓	IPFSa	Ethereum	Not applicable	Not applicable	✓ (Not specified)	✓ (Not specified)	✓ (conceptual token – CID)
2.	[6]	✓	IPFS	Ethereum	Not applicable	✓ (ECCc)	✓ (ECCc)	✓ (MD5b)	✓ (digital token – unspecified)
3.	[34]	N/A	Not specified	Own platform	Not applicable	✓ (RSAf)	Not applicable	Not applicable	Not applicable
4.	[35]	✓	Not specified	Own platform	✓ (AESd)	✓ (public/private key)	✓ (Tx signature)	Not applicable	Not applicable
5.	[3]	✓	Not specified	Ethereum	✓ (AES)	Not applicable	✓ (ECC)	✓ (MACe)	Not applicable
6.	[36]	N/A	Not stated	Not stated	Not applicable	Not applicable	Not applicable	Not applicable	✓ (NFTs – conceptual)
7.	[4]	✓	IPFS (implied)	Ethereum	Not applicable	✓ (ECC)	✓ (ECC)	Not applicable	✓ (NFTs – implied)

[a]IPFS – InterPlanetary File System, [b]MD5 – Message-Digest Algorithm, [c]ECC – Elliptic Curve Cryptography, [d]AES – Advanced Encryption Standard, [e]MAC – Message Authentication Code, [f]RSA – Rivest-Shamir Adleman

V. TAXONOMY OF BLOCKCHAIN-BASED E-BOOK PUBLISHING MODELS

To systematically evaluate the diversity of blockchain applications in e-book publishing, this study develops a taxonomy framework that classifies prior works across six dimensions: platform, storage, smart contract usage, cryptographic algorithm, tokenization, and actors involved. This taxonomy serves two (2) purposes: first, to consolidate scattered findings into a structured framework; and second, to highlight the positioning of the proposed conceptual model within the broader research landscape.

As summarized in Table III, most models rely on the Ethereum blockchain, while some use custom-built or other blockchain platforms. Storage mechanisms vary from centralized cloud storage to decentralized solutions such as IPFS. For instance, [38] proposed a cross-chain copyright protection scheme in which storage and verification are distributed across multiple blockchains, supported by arbitration and supervision entities. This approach demonstrates how storage mechanisms can be extended beyond single-chain IPFS or cloud solutions. Smart contracts are primarily used in these solutions. The cryptographic algorithm implemented includes symmetric encryption (e.g., AES), asymmetric encryption (e.g., RSA, ECC), digital signatures (e.g., EDCSA, transaction-based signatures), and hashing (e.g., MD5, SHA). Similarly, the SecureRights framework by [39] enhances authenticity and digital rights management (DRM) by combining a blockchain smart contract with cryptographic watermarking, fingerprinting, and IPFS. This demonstrates how cryptographic techniques can be extended beyond conventional encryption and hashing to support end-to-end protection of digital content. Tokenization ranges from none to fungible tokens to NFTs used to represent ownership or incorporate publishers as intermediaries in the distribution process.

TABLE III. TAXONOMY OF BLOCKCHAIN-BASED E-BOOK PUBLISHING MODELS

Dimension	Categories / Options	Examples in Literature
Platform	Ethereum, Own Blockchain, Other (e.g., Hyperledger)	[4], [5], [6], [35]
Storage	On-chain, Cloud-based, IPFS / Decentralized	[6] – IPFS; [35] – NFT + external storage
Smart Contract	Royalty automation, Authenticity verification, Marketplace	[5] – Royalty automation; [3] – Marketplace
Cryptographic Algorithm	AES, RSA, ECC, Hashing (MD5, Keccak256, SHA)	[34] – RSA; [3] – AES + MAC
Tokenization	None, Fungible Token (ERC-20), Non-Fungible Token (ERC-721)	[36] and [4] – NFT;
Actors Involved	Author + Customer, Author + Publisher + Customer	Most self-publishing models involve Author + Customer; the Proposed model includes the publisher.

The taxonomy framework is illustrated in Fig. 5, which visualizes the classification across six dimensions. The diagram highlights how blockchain-based models can be categorized based on their technical implementation and functional scope. Together, Table III and Fig. 5 provide a comprehensive reference structure for analyzing similarities and differences among blockchain-based approaches to e-book publishing.

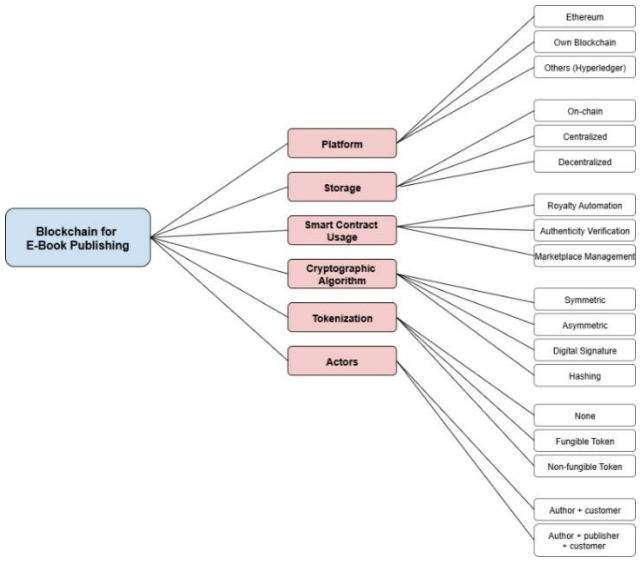


Fig. 5. Taxonomy of blockchain-based e-book publishing models.

A key insight from this taxonomy is that existing works are often fragmented, with most solutions addressing only one or two dimensions at a time. For example, [6] focus primarily on authenticity and storage (IPFS), while [36] highlight copyright protection via NFTs, but overlook encryption and publisher involvement. Similarly, [34] implements asymmetric cryptography but lacks integration with decentralized storage. Few models integrate multi-actor involvement, hybrid cryptography, and automated royalty payments within the same framework. This observation underlines the distinctiveness of the proposed conceptual model, which consolidates these features into a unified approach designed for practical application in e-book publishing.

VI. PROPOSED SOLUTIONS

This section presents a proposed conceptual model of blockchain-based e-book publishing using a smart contract on the Ethereum blockchain. The model aims to provide transparent transactions, verify the authenticity of e-books, and protect authors' copyright by delivering fair royalties and payments to publishers. Fig. 6 provides a general overview of the Blockchain-based E-Book Publishing conceptual model and the interactions among the parties involved with a smart contract on the Ethereum blockchain. This proposed model cannot be applied to self-publishing authors and the marketplace, as there is no direct participation of the publisher and customer. It involves four (4) main actors: the author(s), the publisher, the customer, and the smart contract.

The proposed model applies to a single publisher, which could facilitate e-book marketing and distribution through the publisher's sales platform. The business procedures and internal system processes will take place as follows:

- The publisher will upload an encrypted, signed e-book to the external storage IPFS.
- The details of the e-book, such as the title, price, author(s), and book category, will be added to the system. The contract will be initiated, but it is noted that it is not yet official.
- Hash code generated from IPFS will be stored in the blockchain.
- The author will be notified about the new contract and receive a request for e-book hash verification.
- The author will verify the new contract and the hash code to validate the authenticity of the e-book uploaded into the system. After the author's verification, the new contract will be officially created. The e-book will now be visible for sale.
- Customers will request an e-book by sending a purchase request to the system.
- The request will be notified and reviewed. Upon acceptance, the smart contract will generate the payment request function. The customer will be authorized first to ensure that they are a legitimate buyer.
- The customer should deposit the payment.
- The payment will be validated. The transaction will be aborted if the customer fails to pay within the allocated time.
- Upon successful payment, the smart contract will request the user to verify the e-book hash value for authentication.
- The hash verification by the customer. This ensures the authenticity of the e-book and that the person verifying the hash is the actual purchaser.

- A download link will be provided if the hash verification is successful.
- The customer will download or may access the e-book from IPFS.
- Successful download/access notifications will be sent to the smart contract.
- After a successful download, the smart contract will notify the author and publisher of the sales.
- Royalty payment to the author(s) and sales payment to the publisher will be delivered according to the pre-determined rules set up at the beginning of the smart contract's creation.
- Alternatively, the author(s) and publisher can also request the sales statement from the system.

To further illustrate how the proposed model can be applied in the e-book publishing environment, consider a real e-book purchase by a customer. When the customer initiates the purchase, the smart contract triggers an authenticity verification by comparing the e-book's IPFS hash with the metadata hash stored on the blockchain. If the values match, the system grants access to the encrypted file and automatically releases royalty payments to the author and publisher in accordance with predefined rules. This IPFS-anchored hash verification pattern has been widely adopted in blockchain-based digital publishing applications, indicating its practical feasibility [6], [40], [41], [42].

The following section discusses the theoretical and practical implications of this study, situates the model within the existing body of knowledge, and highlights its potential relevance to stakeholders in the e-book publishing industry.

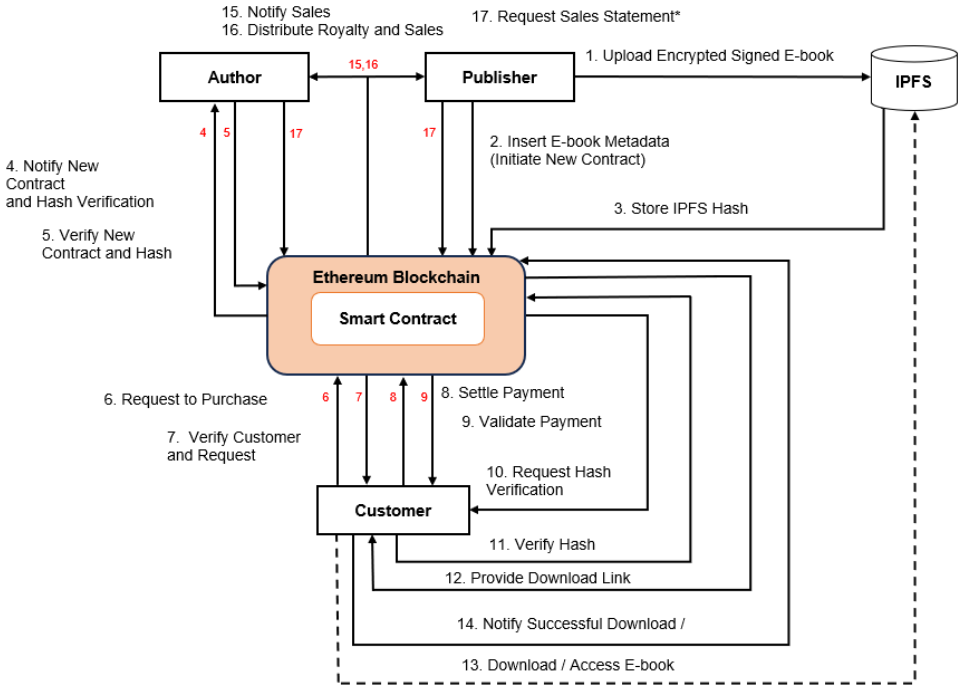


Fig. 6. The conceptual model of blockchain-based e-book publishing.

VII. Discussion

The proposed conceptual model addresses the limitations observed in prior work on the blockchain-based e-book publishing approach. Tables I and II, together with the proposed taxonomy, give a clear overview of how the proposed model is different from previous blockchain-based e-book publishing solutions. While past studies mostly focused on one or two specific functions, such as transaction transparency or authenticity, this study combines multiple essential elements, including storage, cryptographic methods, and actor involvement, into one complete and realistic workflow.

Beyond outlining the model, it is necessary to consider the broader implications for theory and practice, as well as the limitations that will shape future work.

A. Research Implications

This study contributes to the theoretical and practical insights into the domain of blockchain-based e-book publishing, as described below:

1) Theoretical implications: The paper advances blockchain research by locating e-book publishing as a distinct case within digital content publication, where interaction between the authors, publisher, and customers is more complex compared to music and image publishing. By integrating an Ethereum smart contract, IPFS, and a cryptographic algorithm, this model demonstrates how multi-actor relationships can be supported transparently and securely. This study develops a taxonomy for a blockchain-based model of e-book publishing,

classifying it across six dimensions. This taxonomy provides a structured framework for synthesizing prior works and systematically identifying research gaps, thereby serving as a reference for future studies. This aligns with broader blockchain research by [21], which emphasizes the importance of structured frameworks and systematic taxonomies in advancing the blockchain research agenda. The comparative review of seven (7) models further strengthens theoretical understanding by highlighting persistent issues around transparent transactions, e-book authenticity, and copyright protection.

2) Practical implications: For publishers, this model offers a pathway to improve the copyright protection for authors by enforcing automated royalty distribution, thus reducing disputes and building trust in e-book publishing transactions. The model enhances copyright protection through two mechanisms: 1) each e-book will be authorized by the author using hash verification, and 2) royalty payments are executed automatically through smart contracts, ensuring transparency and reducing the sales record manipulation. Author benefit from enhanced visibility and control over sales records, addressing their long-standing concern about the transparency of the royalties. Customers gain assurance that the purchased e-books are authentic and verified. For policymakers, the model demonstrates how blockchain can be applied beyond cryptocurrency to support intellectual property protection. The taxonomy also has practical implications for decision-makers, enabling stakeholders to evaluate existing blockchain solutions based on platform, storage, smart contract usage, cryptographic algorithm, tokenization, and the actors involved. Unlike the previous blockchain model for the digital content that focused on a single function, such as authenticity or transparency, the proposed model's significance lies in its ability to integrate multiple critical e-book publishing functions into a unified and automated workflow.

B. Research Limitations

Although this study contributes to a conceptual blockchain-based model for e-book publishing, several limitations must be acknowledged. First, the research is still conceptual in nature and has not yet been validated through a prototype implementation or testing. In addition, the model is currently limited to a single-publisher scenario. Thus, the model does not apply to a self-publishing and marketplace developed by a third party.

While blockchain offers transparency, e-book authenticity, and robust copyright protection in e-book publishing, its implementation in the real world faces several key challenges. One of the main issues is scalability. Transaction processing time and gas fees can increase significantly when transaction volume is high, potentially delaying e-book verification and payment confirmation. Another key limitation is user adoption. Publisher, author, and customer may not be familiar with blockchain operations, which could slow adoption, especially when the transaction process involves additional steps such as e-wallet creation or gas fee payments. Finally, the integration with existing e-book publishing infrastructure requires the publisher to adjust its internal operations and administrative organization, which can increase the operational costs and complexity. These factors need to be considered to ensure the system's effectiveness, cost efficiency, and usability in practical development.

VIII. Conclusion and Future Works

This paper examined existing blockchain-based approaches to e-book publishing. Through a comparative analysis of seven (7) related works, the persistent gaps in ensuring transparent transactions, e-book authenticity, and copyright protection for the authors are identified. To address these gaps, a conceptual blockchain-based model was proposed that leverages Ethereum smart contract, IPFS, and cryptographic algorithms to provide transparency, authenticity verification, and fair royalty distribution.

The study makes four (4) key contributions: 1) a conceptual model that integrates transparency, authenticity, and copyright protection in a unified model, 2) explicit inclusion of multi-actor involvement – author, publisher, and customer, reflecting real-world e-book publishing process, 3) comparative analysis of prior models to identify limitations and position of the proposed solution, and 4) the development of the taxonomy that establishes a theoretical reference for future study. Together, these contributions provide both theoretical value by advancing blockchain research in e-book publishing and practical significance by offering a foundation for the publishers and authors to adopt with secure e-book distribution mechanisms.

While no prototype or simulation was conducted at this stage, it sets the groundwork for future development. Planned work includes implementing a prototype in Solidity and Ethereum with IPFS storage, evaluating its feasibility with experts, and conducting experimental performance validation in a real e-book publishing scenario. Expanding the model to multi-author environments and integration with tokenization mechanisms, such as NFTs, is also part of the future directions.

From a feasibility perspective, implementing this model on the Ethereum network needs to take into consideration gas fees and scalability constraints. Transaction costs can increase during peak network usage, affecting the user experience. However, using a second layer (Layer-2) solution such as Polygon or Arbitrum can help reduce costs and speed up transactions [43]. Future studies could explore these combinations to improve system performance. By pursuing these next steps, the proposed model can evolve into a validated solution that enhances trust, fairness, and security in the e-book publishing ecosystem.

References

[1] A. Savelyev, "Copyright in the blockchain era: Promises and challenges," Computer Law and Security Review, vol. 34, no. 3, pp. 550–561, Jun. 2018, doi: 10.1016/j.clsr.2017.11.008.

[2] L. Manley and R. P. Holley, "History of the Ebook: The Changing Face of Books," Technical Services Quarterly, vol. 29, no. 4, pp. 292–311, Oct. 2012, doi: 10.1080/07317131.2012.705731.

[3] J. Chi, J. Lee, N. Kim, J. Choi, and S. Park, "Secure and Reliable Blockchain-based eBook Transaction System for Self-published eBook Trading," PLoS One, vol. 15, no. 2, Feb. 2020, doi: 10.1371/journal.pone.0228418.

[4] M. I. S. Ayasy and A. M. Barmawi, "Protecting Author Royalty of Digital Assets Using Blockchain and Elliptic Curve Cryptography," Institute of

Electrical and Electronics Engineers (IEEE), Jan. 2022, pp. 86–92. doi: 10.1109/gecost55694.2022.10010412.

[5] N. Nizamuddin, H. Hasan, K. Salah, and R. Iqbal, "Blockchain-Based Framework for Protecting Author Royalty of Digital Assets," Arab J Sci Eng, vol. 44, no. 4, pp. 3849–3866, Apr. 2019, doi: 10.1007/s13369-018-03715-4.

[6] N. Nizamuddin, H. R. Hasan, and K. Salah, "IPFS-blockchain-based authenticity of online publications," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2018, pp. 199–212. doi: 10.1007/978-3-319-94478-4_14.

[7] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System." [Online]. Available: www.bitcoin.org

[8] P. Zheng, Z. Jiang, J. Wu, and Z. Zheng, "Blockchain-Based Decentralized Application: A Survey," IEEE Open Journal of the Computer Society, vol. 4, pp. 121–133, 2023, doi: 10.1109/OJCS.2023.3251854.

[9] C. Tuteja, N. Saxena, P. Johri, and V. R. Vadi, "Blockchain Technology: A case study of its Decentralized Use," in 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, COM-IT-CON 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 771–775. doi: 10.1109/COM-IT-CON54601.2022.9850511.

[10] A. H. Mohammed, A. A. Abdulateef, and I. A. Abdulateef, "Hyperledger, Ethereum and Blockchain Technology: A Short Overview," in HORA 2021 - 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications, Proceedings, Institute of Electrical and Electronics Engineers Inc., Jun. 2021. doi: 10.1109/HORA52670.2021.9461294.

[11] N. Six, N. Herbaut, and C. Salinesi, "Blockchain Software Patterns for the Design of Decentralized Applications: A Systematic Literature Review," Blockchain: Research and Applications, vol. 3, no. 2, Jun. 2022, doi: 10.1016/j.bcra.2022.100061.

[12] T. Ali Syed, A. Alzahrani, S. Jan, M. S. Siddiqui, A. Nadeem, and T. Alghamdi, "A Comparative Analysis of Blockchain Architecture and its Applications: Problems and Recommendations," IEEE Access, vol. 7, pp. 176838–176869, 2019, doi: 10.1109/ACCESS.2019.2957660.

[13] C. Armstrong, "Books in a virtual world: The evolution of the e-book and its lexicon," Sep. 2008. doi: 10.1177/0961000608092554.

[14] T. Hewa, M. Ylianttila, and M. Liyanage, "Survey on blockchain based smart contracts: Applications, opportunities and challenges," Mar. 01, 2021, Academic Press. doi: 10.1016/j.jnca.2020.102857.

[15] M. Vassiliou and J. Rowley, "Progressing the definition of 'e-book,'" Library Hi Tech, vol. 26, no. 3, pp. 355–368, 2008, doi: 10.1108/07378830810903292.

[16] C. C. Lin, W. C. Chiou, and S. S. Huang, "The challenges facing e-book publishing industry in Taiwan," in Procedia Computer Science, Elsevier B.V., 2013, pp. 282–289. doi: 10.1016/j.procs.2013.05.037.

[17] An Alliance of Independent Authors, "Authors and the Blockchain Towards a Creator Centered Business Model," London. Accessed: Oct. 17, 2024. [Online]. Available: https://www.allianceindependentauthors.org/wordpress/wp-content/uploads/2017/07/Authors-and-the-Blockchain_Towards-a-Creator-Centered-Business-Model.pdf

[18] H. Guo and X. Yu, "A survey on blockchain technology and its security," Blockchain: Research and Applications, vol. 3, no. 2, Jun. 2022, doi: 10.1016/j.bcra.2022.100067.

[19] P. De Filippi, M. Mannan, and W. Reijers, "Blockchain as a confidence machine: The problem of trust & challenges of governance," Technol Soc, vol. 62, Aug. 2020, doi: 10.1016/j.techsoc.2020.101284.

[20] C. H. V. N. U. Bharathi Murthy, M. L. Shri, S. Kadry, and S. Lim, "Blockchain based cloud computing: Architecture and research challenges," IEEE Access, vol. 8, pp. 205190–205205, 2020, doi: 10.1109/ACCESS.2020.3036812.

[21] A. Shukla, P. Jirli, A. Mishra, and A. K. Singh, "An overview of blockchain research and future agenda: Insights from structural topic modeling," Journal of Innovation and Knowledge, vol. 9, no. 4, Oct. 2024, doi: 10.1016/j.jik.2024.100605.

[22] M. Kassab, "Exploring Non-Functional Requirements for Blockchain-Oriented Systems," in Proceedings of the IEEE International Conference on Requirements Engineering, IEEE Computer Society, Sep. 2021, pp. 216–219. doi: 10.1109/REW53955.2021.00040.

[23] "Decentralized Storage | ethereum.org." Accessed: Oct. 17, 2024. [Online]. Available: https://ethereum.org/en/developers/docs/storage/

[24] P. A. Lobo and V. Sarasvathi, "Distributed File Storage Model using IPFS and Blockchain," in 2021 2nd Global Conference for Advancement in Technology, GCAT 2021, Institute of Electrical and Electronics Engineers Inc., Oct. 2021. doi: 10.1109/GCAT52182.2021.9587537.

[25] T. Wang, S. Chang Liew, and S. Zhang, "PubChain: A decentralized open-access publication platform with participants incentivized by blockchain technology," in 2020 International Symposium on Networks, Computers and Communications, ISNCC 2020, Institute of Electrical and Electronics Engineers Inc., Oct. 2020. doi: 10.1109/ISNCC49221.2020.9297213.

[26] D. Sharma and M. Saxena, "Different Cryptographic Hash Functions for Security in the Blockchain," in 2023 International Conference on Data Science and Network Security, ICDSNS 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICDSNS58469.2023.10245326.

[27] S. Ahmad, S. K. Arya, S. Gupta, P. Singh, and S. K. Dwivedi, "Study of Cryptographic Techniques Adopted in Blockchain," in 4th International Conference on Intelligent Engineering and Management, ICIEM 2023, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICIEM59379.2023.10166591.

[28] K. Saini, S. Sharma, and U. Sarkar, "Blockchain and Cryptography," in Proceedings - 2022 4th International Conference on Advances in Computing, Communication Control and Networking, ICAC3N 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 1863–1868. doi: 10.1109/ICAC3N56670.2022.10074345.

[29] R. Sharma, S. Pawar, S. Gurav, and P. Bhavathankar, "A unique approach towards image publication and provenance using blockchain," in Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020, Institute of Electrical and Electronics Engineers Inc., Aug. 2020, pp. 311–314. doi: 10.1109/ICSSIT48917.2020.9214203.

[30] Y. Jiang and J. Zhou, "Digital Music Copyright Protection System Based on Blockchain," in 2022 4th International Academic Exchange Conference on Science and Technology Innovation, IAECST 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 962–966. doi: 10.1109/IAECST57965.2022.10062214.

[31] E. B. E. Correa, V. De B. Nascimento, and A. J. G. Abelem, "DASP: Distributed and Autonomic Scientific Publisher Proposal for Editorial Process Management on Permissioned Blockchain," in 2021 3rd Conference on Blockchain Research and Applications for Innovative Networks and Services, BRAINS 2021, Institute of Electrical and Electronics Engineers Inc., Sep. 2021, pp. 25–26. doi: 10.1109/BRAINS52497.2021.9569805.

[32] J. Zheng, S. Teng, P. Li, W. Ou, D. Zhou, and J. Ye, "A Novel Video Copyright Protection Scheme Based on Blockchain and Double Watermarking," Security and Communication Networks, vol. 2021, 2021, doi: 10.1155/2021/6493306.

[33] M. Maheswari, S. Prasath, R. Rajesh, and D. Ramalakshmi, "Blockchain-based Access Control Model for Student Academic Record with Authentication," in Proceedings of the 7th International Conference on Intelligent Computing and Control Systems, ICICCS 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1411–1414. doi: 10.1109/ICICCS56967.2023.10142704.

[34] J. Putsorn, S. Nontree, and T. Chomsiri, "JS Digital Assets Trading System," in 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON), IEEE, 2019. Accessed: Oct. 18, 2022.[Online]. Available: https://ieeexplore.ieee.org/document/8692301

[35] T. Chomsiri and D. Pansa, "JSP Digital Asset Trading System," in 2019 23rd International Computer Science and Engineering Conference (ICSEC), 2019. doi: 10.1109/ICSEC47112.2019.8974847.

[36] M. Rahrouh, W. Alayash, and M. Ghanem, "The Potential Application Of NFT in the Publishing Industry; Opportunities and Challenges," in 2022 International Arab Conference on Information Technology (ACIT), IEEE, Nov. 2022, pp. 1–5. doi: 10.1109/ACIT57182.2022.9994159.

[37] "Deploying a smart contract to a channel — Hyperledger Fabric Docs main documentation." Accessed: Oct. 17, 2024. [Online]. Available: https://hyperledgerfabric.readthedocs.io/en/latest/deploy_chaincode.html

[38] R. Xie and M. Tang, "A digital resource copyright protection scheme based on blockchain cross-chain technology," Heliyon, vol. 10, no. 17, Sep. 2024, doi: 10.1016/j.heliyon.2024.e36830.

[39] T. Madushanka, D. S. Kumara, and A. A. Rathnaweera, "SecureRights: A Blockchain-Powered Trusted DRM Framework for Robust Protection and Asserting Digital Rights," 2024.

[40] J. Anthal, S. Choudhary, and R. Shettiyar, "Decentralizing File Sharing: The Potential of Blockchain and IPFS," in 2023 International Conference on Advancement in Computation and Computer Technologies, InCACCT 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 773–777. doi: 10.1109/InCACCT57535.2023.10141817. [41]M. M. Arer, P. M. Dhulavvagol, and S. G. Totad, "Efficient Big Data Storage and Retrieval in Distributed Architecture using Blockchain and IPFS," in 2022 IEEE 7th International conference for Convergence in Technology, I2CT 2022, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/I2CT54291.2022.9824566.

[41] H. S. Huang, T. S. Chang, and J. Y. Wu, "A secure file sharing system based on IPFS and blockchain," in ACM International Conference Proceeding Series, Association for Computing Machinery, Jul. 2020, pp. 96–100. doi: 10.1145/3409934.3409948.

[42] M. Mandal, M. S. Chishti, and A. Banerjee, "Investigating Layer-2 Scalability Solutions for Blockchain Applications," in Proceedings - 2023 IEEE International Conference on High Performance Computing and Communications, Data Science and Systems, Smart City and Dependability in Sensor, Cloud and Big Data Systems and Application, HPCC/DSS/SmartCity/DependSys 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 710–717. doi: 10.1109/HPCC-DSS-SmartCity-DependSys60770.2023.00101.

Adaptive Hybrid Deep Learning with Recursive Feature Elimination for Physical Violence Detection

Sukmawati Anggraeni Putri[1*], Duwi Cahya Putri Buani[2], Achmad Rifa'i[3], Imam Nawawi[4]
Department of Information System, Universitas Nusa Mandiri, Jakarta, Indonesia [1]
Department of Informatic, Universitas Nusa Mandiri, Jakarta, Indonesia [2, 3]
Department of Information System, Universitas Bina Sarana Informatika, Jakarta, Indonesia [4]

***Abstract*—Physical violence among students remains a persistent issue that often goes undetected, especially in school environments without intelligent real-time monitoring systems. Such incidents pose serious risks to student safety and hinder the creation of a secure learning atmosphere. This study aims to develop an adaptive visual-based system for detecting physical violence in educational settings using a deep learning approach. A hybrid architecture was designed by integrating VGG19 for spatial feature extraction and Bidirectional Long Short-Term Memory (BiLSTM) for temporal sequence analysis. To enhance model interpretability and reduce redundancy, Recursive Feature Elimination (RFE) was employed to eliminate irrelevant features and improve overall learning efficiency. The proposed system effectively captures both spatial and temporal cues from classroom surveillance videos, enabling more accurate classification of violent and non-violent behaviors. The model was trained and tested on benchmark datasets containing diverse video samples and achieved an accuracy of 92.4%, outperforming standalone CNN and LSTM models. The integration of RFE contributed to a more compact and computationally efficient framework. This study demonstrates the potential of hybrid deep learning and feature optimization for real-time violence detection, contributing to the advancement of visual intelligence and Educational AI for safer, data-driven learning environments.**

***Keywords*—Violence detection; deep learning; VGG19; BiLSTM; RFE; Educational AI**

I. Introduction

Acts of physical aggression in educational settings seriously hinder the psychological and emotional well-being of learners. Even though multiple measures have been introduced to ensure a secure learning atmosphere, such behaviours still occur discreetly, especially in classes with limited monitoring [1]. Without a smart monitoring system, numerous cases go unnoticed in real time, which heightens the possibility of victims experiencing lasting psychological harm. These conditions highlight the urgent need for an adaptive and intelligent system capable of detecting violent actions automatically and accurately in school environments.

Prior studies have investigated the application of machine learning and computer vision for violence detection in video surveillance [2]. Nevertheless, such techniques remain limited, as conventional Convolutional Neural Network (CNN) models extract only spatial information from single frames [3][4][5]. These models neglect the temporal dimension, reducing their effectiveness in recognising repetitive or sequential behavioural patterns over time [3][6]. In contrast, the Long Short-Term Memory (LSTM) approach [7] is able to process temporal dependencies but often depends heavily on the quality of input features and struggles to represent complex visual dynamics [8]. Furthermore, previous studies have generally overlooked feature selection, resulting in increased computational costs and a higher tendency toward overfitting [9].

To address these limitations, this study introduces an adaptive hybrid deep learning framework that integrates both spatial [10] and temporal information simultaneously [11][12]. The proposed framework employs VGG19[13] as its backbone to extract spatial features from individual video frames. In this stage, VGG19 generates feature vectors rather than performing classification, which allows the network to serve purely as a feature extractor. The choice of VGG19 [14] is based on its robust yet compact architecture and its proven effectiveness in object and texture recognition tasks[15][16]. The extracted features are then processed by a Bidirectional Long Short-Term Memory (BiLSTM) network [17][18], enabling the model to capture temporal dependencies in both forward and backward directions [19], thereby improving its ability to recognise patterns of violent behavior [20].

Further enhancing model efficiency and accuracy, the Recursive Feature Elimination (RFE) [21] method is applied to iteratively remove irrelevant or redundant attributes. This process ensures that only the most informative features are retained, effectively reducing model complexity and minimising overfitting risks [22]. Such an approach is particularly beneficial for video-based data[23], which inherently involves high-dimensional spatial–temporal representations [24][25]. By combining spatial and temporal feature analysis with systematic feature selection, the proposed hybrid model offers a comprehensive and efficient solution for detecting violent incidents in educational environments [26]. The contributions of this research are expected to advance the development of artificial intelligence-based educational safety systems that support human-centred, data-driven, and adaptive learning ecosystems in line with current edutech innovations.

II. Related Work

Research on violence detection through video surveillance has expanded significantly, particularly in the domains of public safety and education. One of the initial methods involved the application of Convolutional Neural Networks (CNN) [27][3] to classify violent behaviour using individual frames. For example, Sakhthivinayagam et al. [28] introduced a fast

*Corresponding author.

violence detection framework based on CNN. Nevertheless, this strategy is limited to static image analysis and fails to capture motion dynamics.

In response to these challenges, Sharma et al. [6] proposed a real-time violence detection approach employing a CNN integrated with surveillance systems. The framework could quickly identify violent events but struggled in cases where actions were slow or not overtly visible, suggesting that temporal modelling is essential in detecting aggressive conduct. To address this, Ullah et al. [29] integrated CNN with LSTM, enabling the model to capture both spatial and temporal information [30]. Although their system achieved better results on video datasets, it relied solely on unidirectional LSTM, which processed only past frames and excluded future cues, thereby losing contextual completeness [31]. A number of prior studies have explored hybrid strategies to enhance violence detection performance in surveillance videos. A widely adopted approach is the integration of CNN with LSTM, where CNN extracts spatial features and LSTM handles temporal dynamics [15]. Nonetheless, many of these works continue to employ conventional unidirectional LSTM. For instance, Ullah et al. utilized a CNN–LSTM model for violent video classification, but it only captured sequences from past to future, thus limiting the contextual understanding of events [32].

Building on earlier work, Halder et al. [33] employed Bidirectional LSTM (BiLSTM) for violence detection and showed that bidirectional sequence processing substantially enhances accuracy. By analyzing both past and future frames, BiLSTM is capable of capturing the dynamics of violent events across the entire video [18], offering a more comprehensive temporal representation. For spatial feature extraction, VGG19 was selected due to its architectural depth and proven capability in consistently capturing fine-grained spatial details [34]. This network has been widely applied in pattern recognition research and has demonstrated effectiveness in detecting crucial visual cues such as facial expressions, body postures, and interpersonal interactions. Nonetheless, CNN-based models like VGG19 tend to generate a large volume of features, which may lead to overfitting and extended training times [35]. To mitigate this issue, feature selection techniques such as Recursive Feature Elimination (RFE) have been introduced. As reported by Konyo et al. [22], RFE can efficiently retain the most relevant features by iteratively discarding those with minimal contribution to performance. Despite its potential, the integration of RFE into deep learning, particularly within spatial–temporal hybrid architectures, remains underexplored, even though it offers promising improvements in computational efficiency and model interpretability [36].

Accordingly, this study combines three key components: VGG19 for extracting spatial representations, BiLSTM for modelling bidirectional temporal dependencies, and RFE for optimising feature selection into a unified adaptive intelligent framework. This design is anticipated to enhance the accuracy, efficiency, and interpretability of physical violence detection in educational environments [37]. Furthermore, integrating an automated violence detection system with a psychological risk assessment module that accounts for incident frequency and intensity is essential for building a comprehensive solution to safeguard students [38].

III. Proposed Methodology

This research introduces an adaptive intelligent system architecture aimed at detecting physical violence and estimating psychological risk among students. The framework is composed of eight core components: 1) classroom surveillance video as the input; 2) preprocessing involving frame extraction and normalization; 3) spatial representation learning through VGG19; 4) temporal sequence modeling with BiLSTM; 5) feature refinement using Recursive Feature Elimination (RFE); 6) a classification module for violence detection; 7) a module for computing the psychological risk index; and 8) system output in the form of detection summaries and risk alerts. By integrating spatial–temporal analysis with video-based psychosocial risk assessment, this architecture offers a comprehensive solution for safeguarding students (see Fig. 1).

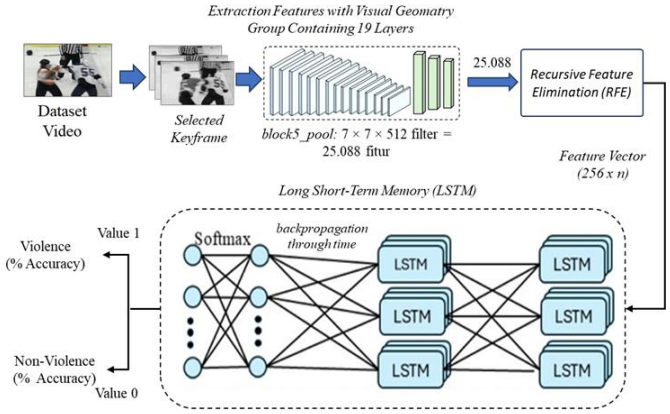


Fig. 1. An overview of the architecture.

A. Dataset

The proposed model was evaluated using three benchmark datasets commonly used for distinguishing between violent and non-violent actions [39][40]. The first is the Hockey Fight Dataset [41], which contains short video clips from ice hockey games, capturing both violent and non-violent interactions. The second is the Violent Flow Dataset [40], consisting of clips extracted from various action-focused movies. The third is the AIRTLab Dataset [42], which comprises video sequences recorded in a simulated environment specifically designed for the development and testing of violence detection systems.

Among these, the Hockey Fight Dataset was selected for training and evaluation of the proposed model due to its balanced composition (500 violent and 500 non-violent clips) [41], the clarity of physical aggression it presents, and its widespread use in prior violence detection research, enabling objective performance comparison. Each clip in the dataset has a resolution of 360 × 240 pixels and a frame rate of 30 fps.

To increase the dataset's relevance in educational settings, selected frames were further analyzed and preprocessed to emulate classroom-like environments while preserving the spatio-temporal features critical for effective violence detection.

B. Feature Extraction Using VGG19

Following the division of video into frames per second (fps), each frame is preprocessed and resized to 224×224 pixels. Spatial representations are then extracted using the VGG19 network [43]. VGG19 was selected due to its consistent performance in visual pattern recognition and its capability to capture high-level spatial features [44]. The architecture has been extensively applied in image classification and transfer learning tasks, particularly in domains characterised by complex visual structures. In this study, the features obtained from the Fully Connected Layer of VGG19 are utilised as inputs for the temporal modelling module [43]. The basic convolution operation is formulated as:

$$F_{i,j}^{(l)} = \sigma\left(\sum_{m,n} K_{m,n}^{(l)} \cdot I_{i+m,j+n}^{(l-1)} + b^{(l)}\right) \quad (1)$$

where, $F_{i,j}^{(l)}$ it is a feature that outputs to the l screen, $K_{m,n}^{(l)}$ it is a convolution kernel, $I_{i+m,j+n}^{(l-1)}$ it is the input from the previous layer, $b^{(l)}$ it is biased, σ it is an activation function, ReLU: $\sigma(x)$ max $(0, x)$.

C. Temporal Analysis Using Bidirectional LSTM

The spatial features extracted from individual video frames are sequentially organised and passed into the BiLSTM network. In contrast to standard LSTM, which processes data in a unidirectional manner from past to future [45], BiLSTM analyses input sequences in both forward and backwards directions, enabling the model to capture a more comprehensive temporal context of violent events [46]. This architecture is particularly effective for modelling behavioural dynamics, as physical violence typically unfolds through a progression of buildup, peak intensity, and subsequent decline within a relatively short time span [47]. The forward-pass operation of the LSTM cell can be formulated as follows:

Input gate:

$$i_t = \sigma(W_i.\left[h_{t-1,}\, x_t\right] + b_i) \quad (2)$$

Forget gate:

$$f_t = \sigma(W_i.\left[h_{t-1,}\, x_t\right] + b_f) \quad (3)$$

Cell State Update:

$$\widetilde{C_t} = \tanh(W_0.\left[h_{t-1,} x_t\right] + b_{0)} \quad (4)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C_t} \quad (5)$$

Output gate:

$$o_t = \sigma(W_o\ .[h_{t-1}, x_t\ ] + b_o) \quad (6)$$

$$h_t = o_t \odot \ \tanh(C_t) \quad (7)$$

For BiLSTM, a two-way LSTM is used:

$$\overrightarrow{h_t}\ (forward\ LSTM) \quad (8)$$

$$\overleftarrow{h_t}\ (backward\ LSTM) \quad (9)$$

Final Output:

$$h_t = \left[\overrightarrow{h_t}\ ;\ \overleftarrow{h_t}\ \right] \quad (10)$$

In this model, x_t denotes the input at time step t (video frame features), while h_t represents the hidden state, and C_t Is the cell state at time, The parameters W and b correspond to trainable weights and biases, The symbo σ refers to the sigmoid activation, $\odot$ It is an element-wise multiplication, tanh It is a hyperbolic tangent activation function, $\vec{h}_t$ is the output of the LSTM forward, $\overleftarrow{h}_t$ Is the output of backward LSTM.

D. Fitur Selection with Recursive Feature Elimination (RFE)

Spatial and temporal feature combinations are often high-dimensional and redundant. To overcome overfitting and speed up training time, the Recursive Feature Elimination (RFE) feature selection technique is used [22]. RFE works by removing features whose contribution to classification performance is considered insignificant [48]. In this study, RFE was applied after initial model training[24]. Features were selected based on their contribution to the validation F1 -score metric [49]. Efficient feature selection plays an important role in reducing computational complexity and improving system interpretability.

If F is a set of features and $f_i \in F$, then:

$$f_{eliminasi} = \arg\min_{f_1}(Score_{model}(F \backslash \{f_i\})) \quad (11)$$

$$Score(x_j) = \left|\frac{\partial \mathcal{L}}{\partial x_j}\right| \quad (12)$$

$$RFE_{DL}(X, y, n) = \arg\min_{S \subseteq X, |S| = n} \mathcal{L}(f_s(X_s), y) \quad (13)$$

where, x_j is the j th feature of the layer before clarification? $\mathcal{L}$ Adalah fungsi loss (e.g., cross-entropy), $\frac{\partial \mathcal{L}}{\partial x_j}$ which is the derivative of the loss function with respect to the feature x_j, which is sensitivity. f_S It is a model built only with a subset of features S, X_S This is input data that only uses features in the subset S, and n is the number of features targeted for retention.

E. Violence Detection Module

The features produced by the BiLSTM-RFE module are subsequently passed into a dense layer with sigmoid activation, which performs binary classification to decide whether the sequence of frames corresponds to violent behavior. The decision threshold is adjusted according to validation outcomes in order to minimize false positives.

The performance of the model is assessed using accuracy [50], precision [51], recall [52], and F1-score [49], following common evaluation metrics in video classification research. This module determines whether a given video clip depicts violent activity [53]. The spatial features extracted by VGG19 and the temporal features captured by BiLSTM are concatenated and forwarded to a dense layer, where the final decision is made through a Softmax activation function.

If h it is the final representation (combination of VGG19 + BiLSTM), then:

$$P(y = k|h) = \frac{e^{W_k^T h + b_k}}{\sum_{j=1}^{K} e^{W_j^T h + b_j}} \quad (14)$$

Prediction:

$$\hat{y} = \arg\max_k P(y = k|h) \quad (15)$$

where, h it is a combined feature vector from BiLSTM (spatial-temporal), W_k is the boot layer output for class k, b_k is the bias for the class k, K is the number of classes (2: affection/disaffection), $P(y = k|h)$ is the probability that the input belongs to class k, and $\hat{y}$ is the final grade prediction.

F. Evaluation and Validation

System performance evaluation [54], [55] is conducted using four key metrics: accuracy, precision, recall, and F1-score, each of which provides a different perspective on system performance.

TABLE I. CONFUSION MATRIX

	Positive Class	Negative Class
Positive Class	True Positive (TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

Table I shows that TP is True Positive (the model correctly predicts violence), TN is True Negative (the model correctly predicts no violence), FP is False Positive (the model incorrectly predicts violence when there is none), and FN is False Negative (the model fails to detect violence when it exists). Accuracy is used to measure the proportion of correct predictions out of all tested data. This metric provides an overview of the system's performance in distinguishing between violent and non-violent incidents, and is formulated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP=FN} \quad (16)$$

However, in unbalanced data conditions, where the number of violent incidents is far fewer than normal incidents, accuracy alone is not sufficient to fairly represent the system's performance. Therefore, precision and recall metrics are used. Precision measures how accurately the system identifies violent incidents, which is the proportion of predictions of violence that are actually violent. The formula is:

$$Precision = \frac{TP}{TP+FP} \quad (17)$$

Meanwhile, the recall shows how well the system captures all incidents of violence that occur. This metric is particularly important in the context of surveillance systems, where false negatives (FN) can have serious consequences:

$$Recall = \frac{TP}{TP-FN} \quad (18)$$

The balance between precision and recall is represented by the F1-score, which is the harmonic mean of the two. The F1-score provides a fairer assessment when there is an imbalance between positive and negative classes.

$$F1 = 2x\frac{Precision\ x\ Recall}{Precision + Recall}$$

The use of these four metrics enables a more comprehensive and fair evaluation of the system, especially in the domain of violence detection, which has high ethical and social implications. This evaluation is also important as a reference in comparing the effectiveness of various proposed deep learning architectures. In addition, a comparative study (ablation study) was conducted to assess the influence of each main component, namely: Without RFE (VGG19 + BiLSTM), Without BiLSTM (VGG19 + Dense), and Without the psychological module. This approach is important for determining the relative contribution of each system component and ensuring that performance improvements are not the result of overfitting or data bias.

IV. DISCUSSION

This study aims to detect violent actions in videos using a deep learning approach that integrates a Convolutional Neural Network (CNN) for spatial feature extraction and a Long Short-Term Memory (LSTM) for temporal sequence modeling. Recursive Feature Elimination (RFE) is used as the feature selection method to optimize model performance.

Each video is processed by dividing its content into a number of key frames taken at specific time intervals to effectively represent visual information. Spatial feature extraction is performed using the VGG19 CNN architecture, which generates hundreds to thousands of features on each frame or video segment. Due to the high dimension of the features generated, RFE is applied to filter the most relevant features, so that data complexity can be reduced without sacrificing the model's accuracy in classifying violent and non-violent actions.

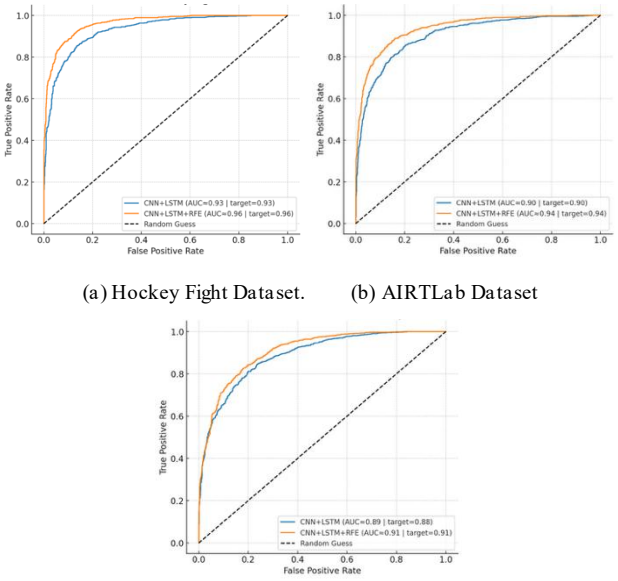
In general, the application of RFE can reduce feature dimensions by 30 to 50%, depending on the configuration used. This reduction has a positive impact on computational efficiency, with a 20 to 35% increase in the performance of the LSTM model training process. The CNN-LSTM model was trained using a number of violent action video datasets, including Hockey Fight, AIRTLab, and Violent-Flows. The evaluation process was conducted using the k-fold cross-validation method, along with a data distribution scheme of 80% for training and 20% for testing.

To evaluate the model's performance in classification, a confusion matrix is used to provide a detailed overview of the number of correct and incorrect predictions in each class. Through the confusion matrix, it is possible to analyze how well the model recognizes each class, including its ability to avoid classification errors. Table II presents the results of the confusion matrix from the trained model.

TABLE II. RESULTS OF THE CONFUSION MATRIX FROM THE TRAINED MODEL

Dataset	Model	TP	FN	TN	FP
Hockey Fight	CNN+BiLSTM	920	80	915	85
	CNN+BiLSTM+RFE	945	55	935	65
AIRTLab	CNN+BiLSTM	895	105	900	100
	CNN+BiLSTM+RFE	930	70	920	80
Violent-Flows	CNN+BiLSTM	875	125	880	120
	CNN+BiLSTM+RFE	905	95	905	95

Evaluate the model's ability to distinguish between classes as a whole, the Receiver Operating Characteristic (ROC) curve is used. The ROC curve illustrates the trade-off between true positive rate (TPR) and false positive rate (FPR) at various classification thresholds. The closer the curve is to the upper left corner, the better the model performance. Fig. 2 shows the ROC curve of the trained model.



(a) Hockey Fight Dataset. (b) AIRTLab Dataset

(c) Violent Flow Dataset

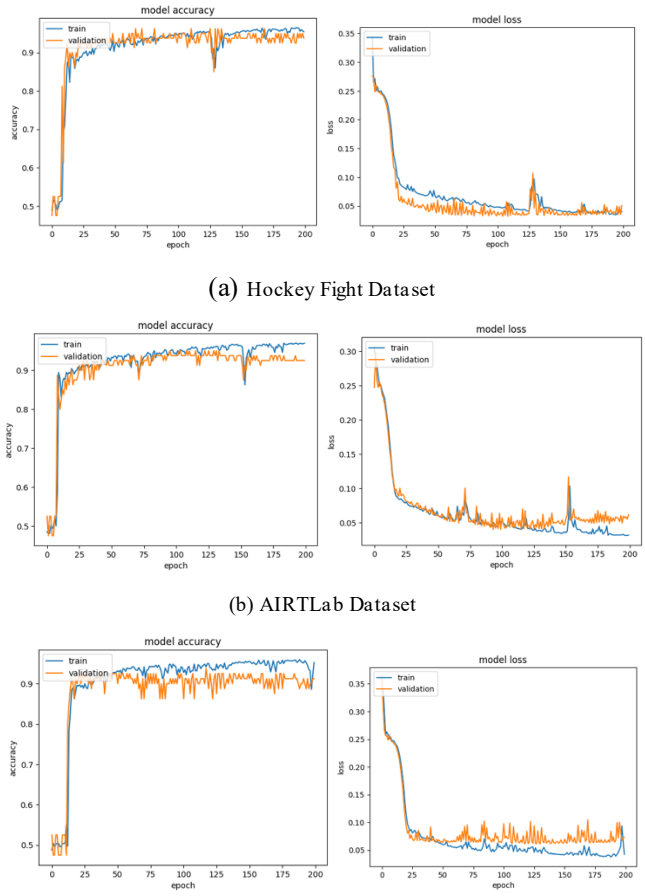
Fig. 2. Receiver Operating Characteristic (ROC) Curve for each Dataset.

Model performance was evaluated using various measurement metrics, including accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) based on the ROC curve. Table III shows the results of model performance evaluation per dataset.

TABLE III. PERCENTAGE OF PERFORMANCE MODEL

Dataset	Model	Accuracy	Precission	Recall	F1-Score
Hockey Fight	CNN+BiLSTM	0.918	0.915	0.920	0.917
	CNN+BiLSTM + RFE	0.940	0.936	0.945	0.941
AIRTLab	CNN+BiLSTM	0.898	0.899	0.895	0.897
	CNN+BiLSTM + RFE	0.925	0.921	0.930	0.925
Violent-Flows	CNN+BiLSTM	0.878	0.879	0.875	0.877
	CNN+BiLSTM + RFE	0.905	0.905	0.905	0.905

Monitor model performance during the training process, model accuracy and model loss graphs are used. The accuracy graph shows the progress of the model's ability to classify correctly, while the loss graph illustrates the amount of error produced by the model. These graphs provide an overview of whether the model is underfitting, overfitting, or has been trained properly. Fig. 3 illustrates the accuracy and loss graphs from the model training process.



(a) Hockey Fight Dataset

(b) AIRTLab Dataset

(c) Violent Flow Dataset

Fig. 3. Performance evaluation results of the CNN + BiLSTM + RFE model on the datasets: (a) Hockey Fight, (b) AIRTLab-2000, (c) Violent Flow.

The three figures described previously show that the application of Recursive Feature Elimination (RFE) has a positive impact in increasing accuracy on the three data sets: Hockey Fight, AIRLab, and Violence Flow. For the Hockey Fight dataset, both training and validation accuracy rose sharply from approximately 0.5 to above 0.9. Training accuracy stabilized between 0.95 and 0.97, while validation accuracy fluctuated slightly within the 0.90–0.95 range. Since validation remained consistent, there was no indication of severe overfitting, making early stopping unnecessary. In the AIRTLab dataset, performance improved rapidly during the early stages, with training accuracy approaching 0.98 and validation accuracy leveling off at 0.93–0.95. A minor drop in validation performance between epochs 120 and 160 suggested mild overfitting, making early stopping around epoch 100–120 a more efficient option. In the Violence Flow dataset, accuracy also improved initially, but validation accuracy fluctuated more than in the other datasets. Training accuracy climbed to nearly 0.97, while validation stagnated around 0.90–0.93, indicating potential overfitting after epoch 100. The widening gap between training and validation results shows that the model tends to overfit and struggles to generalise effectively. Across datasets, the use of Recursive Feature Elimination (RFE) enhanced performance by filtering out irrelevant features. RFE not only increased accuracy and F1-score but also reduced

computational demands, confirming that many CNN-generated features are redundant and unnecessary for violence detection.

Compared to conventional CNN and LSTM models, the proposed VGG19-BiLSTM–RFE framework achieved superior performance with an overall accuracy of 92.4%, outperforming CNN-LSTM (88.5%). The inclusion of RFE enhanced model interpretability by eliminating redundant features, resulting in faster convergence and reduced overfitting. This improvement demonstrates the model's efficiency in handling the complex spatial–temporal relationships inherent in violence detection tasks.

The CNN heatmap results indicate that frames containing intense physical activities, such as body collisions, play a crucial role in the features retained by RFE. At the same time, the LSTM component effectively identifies temporal patterns in violent events, including abrupt and repetitive movements. Incorporating Recursive Feature Elimination (RFE) into the CNN-LSTM pipeline has proven successful in reducing feature dimensionality and computational costs, while also enhancing accuracy, precision, and reliability in violence detection by retaining only the most discriminative spatial features. When applied within the CNN-BiLSTM framework, RFE further improved classification performance by increasing accuracy and F1-scores, while removing redundant features without degrading the model's capability. The heatmap visualisations also confirm that frames with strong physical interactions dominate the important feature set, while BiLSTM captures the bidirectional temporal dynamics characteristic of violent actions. Nonetheless, misclassifications persist in scenarios involving non-violent videos with rapid movement or subtle forms of aggression, suggesting the need for integrating additional temporal descriptors such as optical flow or attention mechanisms. Overall, these findings establish RFE as an efficient and effective feature selection strategy for deep learning–based video analysis, particularly for tackling complex event detection tasks like violence recognition.

V. Conclusion

This study introduced an adaptive hybrid deep learning framework that combines VGG19 for spatial feature extraction, BiLSTM for temporal modeling, and Recursive Feature Elimination (RFE) for feature optimization in video-based violence detection. The model demonstrated high accuracy (92.4%) and outperformed traditional single-model baselines, validating the effectiveness of integrating spatio–temporal learning with feature selection. Evaluation on benchmark datasets confirmed the system's robustness and computational efficiency. Beyond performance metrics, the proposed framework offers practical relevance for real-world applications, particularly in educational settings where early detection of violent behavior is critical. Its interpretable and scalable design supports potential deployment in real-time surveillance systems aimed at enhancing student safety. However, limitations remain in detecting subtle or ambiguous violent actions and in generalizing beyond benchmark datasets. Future work will focus on improving adaptability through attention mechanisms, multimodal fusion (e.g., visual-audio inputs), and real-time system integration. Overall, this research contributes to the development of ethical, interpretable, and data-driven AI solutions for proactive violence prevention and behavioral monitoring in educational environments.

Acknowledgment

The authors gratefully acknowledge the financial support provided by the Directorate of Research and Community Service (DPPM), Ministry of Education, Research, and Technology (Kemendikti-Saintek), through the 2025 research grant, which enabled the completion of this study. The authors also wish to extend their sincere appreciation to all individuals and institutions who contributed to the data collection, research implementation, and preparation of this manuscript, whether directly or indirectly.

References

[1] S. Johansson, E. Myrberg, and A. Toropova, "School bullying: Prevalence and variation in and between school systems in TIMSS 2015," Studies in Educational Evaluation, vol. 74, no. June, p. 101178, Sep. 2022, doi: 10.1016/j.stueduc.2022.101178.

[2] M. Shoaib and N. Sayed, "A Deep Learning Based System for the Detection of Human Violence in Video Data," Traitement du Signal, vol. 38, no. 6, pp. 1623–1635, Dec. 2021, doi: 10.18280/ts.380606.

[3] I.-A. Haiura and A. Iftene, "Detecting Violence in Videos using Convolutional Neural Networks," Procedia Comput Sci, vol. 246, pp. 2497–2506, 2024, doi: 10.1016/j.procs.2024.09.465.

[4] U. Muneer butt, S. Letchmunan, F. Hafinaz Hassan, S. Zia, C. Campus, and P. Anees Baqir, "Detecting Video Surveillance Using VGG19 Convolutional Neural Networks," 2020. [Online]. Available: https://sites.google.com/view/debadityaroy/datasets

[5] I. A. Haiura and A. Iftene, "Detecting Violence in Videos using Convolutional Neural Networks," Procedia Comput Sci, vol. 246, no. C, pp. 2497–2506, Jan. 2024, doi: 10.1016/J.PROCS.2024.09.465.

[6] S. Sharma, B. Sudharsan, S. Naraharisetti, V. Trehan, and K. Jayavel, "A fully integrated violence detection system using CNN and LSTM," International Journal of Electrical and Computer Engineering, vol. 11, no. 4, pp. 3374–3380, 2021, doi: 10.11591/ijece.v11i4.pp3374-3380.

[7] M. Majd and R. Safabakhsh, "Correlational Convolutional LSTM for human action recognition," Neurocomputing, vol. 396, no. xxxx, pp. 224–229, 2020, doi: 10.1016/j.neucom.2018.10.095.

[8] M. Patel, "Real-Time Violence Detection Using CNN-LSTM," 2021, [Online]. Available: http://arxiv.org/abs/2107.07578

[9] J. Liu, P. Dai, G. Han, and N. Sun, "Combined CNN/RNN video privacy protection evaluation method for monitoring home scene violence," Computers and Electrical Engineering, vol. 106, p. 108614, Mar. 2023, doi: 10.1016/J.COMPELECENG.2023.108614.

[10] G. Kaur and S. Singh, "Revisiting vision-based violence detection in videos: A critical analysis," Neurocomputing, vol. 597, p. 128113, Sep. 2024, doi: 10.1016/J.NEUCOM.2024.128113.

[11] S. A. Arun Akash, R. Sri Skandha Moorthy, K. Esha, and N. Nathiya, "Human Violence Detection Using Deep Learning Techniques," J Phys Conf Ser, vol. 2318, no. 1, 2022, doi: 10.1088/1742-6596/2318/1/012003.

[12] A. Ben Mabrouk and E. Zagrouba, "Spatio-temporal feature using optical flow based distribution for violence detection," Pattern Recognit Lett, vol. 92, pp. 62–67, Jun. 2017, doi: 10.1016/J.PATREC.2017.04.015.

[13] S. Letchmunan, U. M. Butt, F. H. Hassan, S. Zia, and A. Baqir, "Detecting video surveillance using VGG19 convolutional neural networks," International Journal of Advanced Computer Science and Applications, vol. 11, no. 2, pp. 674–682, 2020, doi: 10.14569/ijacsa.2020.0110285.

[14] P. Kuppusamy and V. C. Bharathi, "Human abnormal behavior detection using CNNs in crowded and uncrowded surveillance – A survey," Measurement: Sensors, vol. 24, p. 100510, Dec. 2022, doi: 10.1016/J.MEASEN.2022.100510.

[15] M. Qasim and E. Verdu, "Video anomaly detection system using deep convolutional and recurrent models," Results in Engineering, vol. 18, p. 101026, Jun. 2023, doi: 10.1016/J.RINENG.2023.101026.

[16] F. J. Rendón-Segador, J. A. Álvarez-García, J. L. Salazar-González, and T. Tommasi, "CrimeNet: Neural Structured Learning using Vision Transformer for violence detection," Neural Networks, vol. 161, pp. 318–329, Apr. 2023, doi: 10.1016/J.NEUNET.2023.01.048.

[17] S. Singh, S. Dewangan, G. S. Krishna, V. Tyagi, S. Reddy, and P. R. Medi, "Video Vision Transformers for Violence Detection," 2022, [Online]. Available: http://arxiv.org/abs/2209.03561

[18] N. Amer Hamzah and D. B. N. Dhannoon, "Detecting Arabic sexual harassment using bidirectional long-short-term memory and a temporal convolutional network," Egyptian Informatics Journal, vol. 24, no. 2, pp. 365–373, Jul. 2023, doi: 10.1016/J.EIJ.2023.05.007.

[19] N. Dündar, A. S. Keçeli, A. Kaya, and H. Sever, "A shallow 3D convolutional neural network for violence detection in videos," Egyptian Informatics Journal, vol. 26, p. 100455, Jun. 2024, doi: 10.1016/J.EIJ.2024.100455.

[20] R. Halder and R. Chatterjee, "CNN-BiLSTM Model for Violence Detection in Smart Surveillance," SN Comput Sci, vol. 1, no. 4, 2020, doi: 10.1007/s42979-020-00207-x.

[21] S. Xia and Y. Yang, "A model-free multivariate non-recursive feature elimination for feature selection on high-dimensional complex multiple response data," Inf Sci (N Y), vol. 713, p. 122186, Sep. 2025, doi: 10.1016/J.INS.2025.122186.

[22] O. Kornyo et al., "Botnet attacks classification in AMI networks with recursive feature elimination (RFE) and machine learning algorithms," Comput Secur, vol. 135, p. 103456, Dec. 2023, doi: 10.1016/J.COSE.2023.103456.

[23] 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," Eng Appl Artif Intell, vol. 123, p. 106173, Aug. 2023, doi: 10.1016/J.ENGAPPAI.2023.106173.

[24] S. Kollem, C. Sirigiri, and S. Peddakrishna, "A novel hybrid deep CNN model for breast cancer classification using Lipschitz-based image augmentation and recursive feature elimination," Biomed Signal Process Control, vol. 95, p. 106406, Sep. 2024, doi: 10.1016/J.BSPC.2024.106406.

[25] X. Xu, Z. Liao, and Z. Xu, "Violent Physical Behavior Detection using 3D Spatio-Temporal Convolutional Neural Networks." [Online]. Available: www.ijacsa.thesai.org

[26] D. Sriveni and D. L. R, "An active learning driven deep spatio-textural acoustic feature ensemble assisted learning environment for violence detection in surveillance videos," Engineering Science and Technology, an International Journal, vol. 66, p. 102050, Jun. 2025, doi: 10.1016/J.JESTCH.2025.102050.

[27] J. Liu, P. Dai, G. Han, and N. Sun, "Combined CNN/RNN video privacy protection evaluation method for monitoring home scene violence," Computers and Electrical Engineering, vol. 106, p. 108614, Mar. 2023, doi: 10.1016/J.COMPELECENG.2023.108614.

[28] G. Sakthivinayagam, R. Easawarakumar, A. Arunachalam, and P. M, "Violence Detection System using Convolution Neural Network," International Journal of Electronics and Communication Engineering, vol. 6, no. 2, pp. 5–8, 2019, doi: 10.14445/23488549/ijece-v6i2p102.

[29] W. Ullah et al., "Artificial Intelligence of Things-assisted two-stream neural network for anomaly detection in surveillance Big Video Data," Future Generation Computer Systems, vol. 129, pp. 286–297, Apr. 2022, doi: 10.1016/j.future.2021.10.033.

[30] S. K. Parui, S. K. Biswas, S. Das, M. Chakraborty, and B. Purkayastha, "An Efficient Violence Detection System from Video Clips using ConvLSTM and Keyframe Extraction," 11th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks, IEMECON 2023, no. June, pp. 1–5, 2023, doi: 10.1109/IEMECON56962.2023.10092302.

[31] F. U. M. Ullah, A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Violence detection using spatiotemporal features with 3D convolutional neural network," Sensors, vol. 19, no. 11, pp. 1–15, 2019, doi: 10.3390/s19112472.

[32] N. Mumtaz et al., "An overview of violence detection techniques: current challenges and future directions," Artif Intell Rev, vol. 56, no. 5, pp. 4641–4666, 2023, doi: 10.1007/s10462-022-10285-3.

[33] R. Chatterjee and R. Halder, "Discrete Wavelet Transform for CNN-BiLSTM-based Violence Detection," in International Conference on Emerging Trends and Advances in Electrical Engineering and Renewable Energy (ETAEERE -2020), 2020.

[34] Irfanullah, T. Hussain, A. Iqbal, B. Yang, and A. Hussain, "Real time violence detection in surveillance videos using Convolutional Neural Networks," Multimed Tools Appl, vol. 81, no. 26, pp. 38151–38173, Nov. 2022, doi: 10.1007/s11042-022-13169-4.

[35] Zanzan Lu, X. Xia, H. Wu, and C. Yang, "Violence Detection With Two-Stream Neural Network Based on C3D," International Journal of Cognitive Informatics and Natural Intelligence, vol. 15, no. 4, pp. 1–17, 2021, doi: 10.4018/IJCINI.287601.

[36] M. A. Soeleman, C. Supriyanto, D. P. Prabowo, and P. N. Andono, "Experimental Evaluation of Resampling Algorithms on the Imbalance Violence Video Detection," International Journal of Engineering Trends and Technology, vol. 70, no. 7, pp. 260–268, 2022, doi: 10.14445/22315381/IJETT-V70I7P226.

[37] G. Kaur and S. Singh, "Revisiting vision-based violence detection in videos: A critical analysis," Neurocomputing, vol. 597, p. 128113, Sep. 2024, doi: 10.1016/j.neucom.2024.128113.

[38] M. Edo, V. Oubiña, and M. Svarc, "Machine learning and public policy: Early detection of physical violence against children," Child Youth Serv Rev, vol. 166, p. 107932, Nov. 2024, doi: 10.1016/j.childyouth.2024.107932.

[39] M. Bianculli et al., "A dataset for automatic violence detection in videos," Data Brief, vol. 33, p. 106587, 2020, doi: 10.1016/j.dib.2020.106587.

[40] K. Gkountakos, K. Ioannidis, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, "A crowd analysis framework for detecting violence scenes," ICMR 2020 - Proceedings of the 2020 International Conference on Multimedia Retrieval, no. June, pp. 276–280, 2020, doi: 10.1145/3372278.3390725.

[41] M. Bianculli et al., "A dataset for automatic violence detection in videos," Data Brief, vol. 33, p. 106587, Dec. 2020, doi: 10.1016/J.DIB.2020.106587.

[42] P. Sernani, N. Falcionelli, S. Tomassini, P. Contardo, and A. F. Dragoni, "Deep Learning for Automatic Violence Detection: Tests on the AIRTLab Dataset," IEEE Access, vol. 9, pp. 160580–160595, 2021, doi: 10.1109/ACCESS.2021.3131315.

[43] F. A. Memon, M. H. Memon, I. A. Halepoto, R. Memon, and A. R. Bhangwar, "Action Recognition in videos using VGG19 pre-trained based CNN-RNN Deep Learning Model," VFAST Transactions on Software Engineering, vol. 12, no. 1, pp. 46–57, Mar. 2024, doi: 10.21015/vtse.v12i1.1711.

[44] N. Honarjoo, A. Abdari, and A. Mansouri, "Violence Detection Using One-Dimensional Convolutional Networks," 2021 12th International Conference on Information and Knowledge Technology, IKT 2021, no. December, pp. 188–191, 2021, doi: 10.1109/IKT54664.2021.9685835.

[45] M. A. Soeleman, C. Supriyanto, D. P. Prabowo, and P. N. Andono, "Video Violence Detection Using LSTM and Transformer Networks Through Grid Search-Based Hyperparameters Optimization," International Journal of Safety and Security Engineering, vol. 12, no. 05, pp. 615–622, 2022, doi: 10.18280/ijsse.120510.

[46] T. Haque, F. F. Ahmed, S. M. I. Ahmed, and M. Siam, "Optical Flow based Violence Detection from Video Footage using Hybrid MobileNet and Bi-LSTM," no. September, 2023.

[47] B. Omarov, S. Narynov, Z. Zhumanov, A. Gumar, and M. Khassanova, "State-of-the-art violence detection techniques in video surveillance security systems: A systematic review," PeerJ Comput Sci, vol. 8, pp. 1–41, 2022, doi: 10.7717/PEERJ-CS.920.

[48] N. Mumtaz, N. Ejaz, S. Aladhadh, S. Habib, and M. Y. Lee, "Deep Multi-Scale Features Fusion for Effective Violence Detection and Control Charts Visualization," Sensors, vol. 22, no. 23, pp. 1–15, 2022, doi: 10.3390/s22239383.

[49] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, Jan. 2020, doi: 10.1186/S12864-019-6413-7.

[50] R. Vijeikis, V. Raudonis, and G. Dervinis, "Efficient Violence Detection in Surveillance," Sensors, vol. 22, no. 6, 2022, doi: 10.3390/s22062216.

[51] K. Gkountakos, K. Ioannidis, T. Tsikrika, S. Vrochidis, and I. Kompatsiaris, "Crowd Violence Detection from Video Footage," Proceedings - International Workshop on Content-Based Multimedia Indexing, vol. 2021-June, no. September, 2021, doi: 10.1109/CBMI50038.2021.9461921.

[52] D. Choqueluque-Roman and G. Camara-Chavez, "Weakly Supervised Violence Detection in Surveillance Video," Sensors, vol. 22, no. 12, pp. 1–29, 2022, doi: 10.3390/s22124502.

[53] M. Biswas et al., "State-of-the-Art Violence Detection Techniques: A review," Asian Journal of Research in Computer Science, no. February, pp. 29–42, 2022, doi: 10.9734/ajrcos/2022/v13i130305.

[54] Md. B. Rahman, H. A. Mustafa, and M. D. Hossain, "Towards evaluating robustness of violence detection in videos using cross-domain transferability," Journal of Information Security and Applications, vol. 77, p. 103583, Sep. 2023, doi: 10.1016/j.jisa.2023.103583.

[55] S. A. Putri, "HU Variance Moment Optimizes Keyframe Selection Based on Deep Learning for Violence Detection," Journal of Applied Data Sciences, vol. 6, no. 2, pp. 1088–1101, May 2025, doi: 10.47738/jads.v6i2.648.

Transformative Integration of Machine Learning in Software Applications in Light of Current Software Engineering Practices

Fawzi Abdulaziz Albalooshi
Department of Computer Science, University of Bahrain, Sakhir Campus, Kingdom of Bahrain

***Abstract*—This study critically reviews the transformative integration of machine learning (ML) into software engineering, detailing its evolution from traditional DevOps to MLOps, which has significantly enhanced software development by enabling adaptive and intelligent systems, improving processes, and boosting software quality. Despite these benefits, the integration introduces unique challenges across technical (e.g., model deployment, data quality, scalability), organizational (e.g., collaboration, tool management), and cultural (e.g., resistance to change, skill gaps) domains throughout the software development lifecycle. The review highlights emerging solutions, including robust MLOps practices, microservices architecture, and frameworks like CRISP-DM, DataOps, and Agile ML, which aim to streamline the ML lifecycle and ensure reliability and scalability. Furthermore, it emphasizes the crucial role of security and governance frameworks in protecting against adversarial attacks, maintaining data privacy, and ensuring accountability and compliance, which are essential for building trust and ethical application of ML systems. Ultimately, successful ML integration requires a holistic approach that addresses these multifaceted challenges to optimize ML's impact and drive technological progress and business value.**

Keywords*—Machine learning (ML); software engineering; DevOps; MLOps; ML integration challenges; integrated software development

I. Introduction

Machine learning has profoundly transformed software engineering, leading to more adaptive and intelligent systems and improving development processes and software quality. The significance of this work lies in providing a comprehensive review of this transformative integration, particularly its evolution from traditional DevOps to MLOps, and addressing the complex challenges that arise. This study is crucial for optimizing ML's impact, ensuring the scalability, reliability, and maintainability of ML systems, which is vital for organizations seeking operational efficiency and competitive advantage in an increasingly ML-driven world. While the integration of ML offers substantial benefits, it introduces unique and multifaceted challenges across technical, organizational, and cultural domains throughout the software development lifecycle. The existing literature acknowledges these challenges, but many issues remain unresolved or contentious due to the complexity of ML systems and the evolving nature of software engineering practices. There is a clear gap in a consolidated, critical review that systematically analyzes these integration complexities and explores emerging solutions and best practices to address them effectively. This study aims to fill that gap by providing a structured overview of the current landscape, highlighting where current practices fall short and where further innovation is needed. Incorporating machine learning into software engineering enhances development processes, improves software quality, and allows complex tasks to be framed as learning problems. This approach excels at detecting patterns in large datasets and adapting to changing conditions, proving particularly useful in areas where traditional programming methods fall short [1][2]. However, this integration poses several challenges throughout the software development lifecycle, from requirements engineering to security and operationalization. The complexity of ML systems, combined with evolving software engineering practices, creates a landscape where certain issues remain unresolved or contentious [3][4][5][6][68]. This review argues that integrating ML into software engineering offers significant benefits. However, it also presents unique challenges that require innovative solutions and interdisciplinary collaboration. The review examines ML integration in software applications, emphasizing the transition from DevOps to MLOps. It critically analyzes the challenges in ML integration across technical, organizational, and cultural domains and explores potential solutions. The study aims to provide insights into optimizing ML's impact, ensuring the scalability, reliability, and maintainability of ML systems, ultimately leading to improved operational efficiency and competitive advantage. This study makes several key contributions. It provides a critical review of ML's integration into software engineering, detailing its evolution from DevOps to MLOps. The study systematically highlights the technical, organizational, and cultural challenges encountered during ML integration. It explores emerging solutions, including robust MLOps practices, microservices architecture, and frameworks like CRISP-DM, DataOps, and Agile ML, that aim to streamline the ML lifecycle and ensure reliability and scalability. The review underscores the crucial role of security and governance frameworks in protecting against adversarial attacks, maintaining data privacy, and ensuring accountability and compliance, which are essential for building trust and ethical application of ML systems. These contributions are designed to offer comprehensive insights into optimizing ML's impact, ensuring the scalability, reliability, and maintainability of ML systems, and ultimately driving technological progress and business value.

II. Research Methodology

This study adopts a critical review methodology to comprehensively analyze the transformative integration of machine learning (ML) into software engineering practices. This research design is chosen to provide a structured and in-depth examination of the subject, synthesizing existing knowledge, identifying gaps, and proposing solutions. A critical review allows for a systematic exploration of the evolution from traditional DevOps to MLOps, assessing the benefits and challenges inherent in this transition. The justification for this approach lies in its ability to offer a holistic perspective, crucial for understanding a rapidly evolving and multifaceted domain like ML integration in software engineering.

The research questions guiding this study are formulated to address significant gaps and complexities identified in the current literature regarding ML integration. The primary objectives include: understanding the evolution of ML integration, identifying and analyzing challenges, exploring emerging solutions, and emphasizing security and governance. The methodology involves a thorough review of academic literature, including research papers, conference proceedings, and industry reports related to machine learning, software engineering, DevOps, and MLOps. The selection criteria for literature prioritize relevance to the integration of ML in software development, focusing on studies that discuss challenges, solutions, best practices, and ethical considerations. The collected information is then critically analyzed to synthesize findings, identify recurring themes, and pinpoint areas of consensus and contention. This analytical process allows for the construction of a comprehensive overview that addresses the research questions and contributes to filling the identified literature gaps. The ultimate goal is to provide insights into optimizing ML's impact, ensuring the scalability, reliability, and maintainability of ML systems, and driving technological progress and business value.

III. Software Delivery Pipelines: From DevOps to MLOps

A. Traditional DevOps Continuous Integration (CI)/ Continuous Deployment (CD) Pipelines

Traditional DevOps CI/CD pipelines are essential to contemporary software development, allowing teams to automate and streamline the processes of code integration, testing, and deployment. These pipelines boost efficiency, minimize errors, and ensure the swift delivery of software. The core components of a traditional CI/CD pipeline include CI, Continuous Testing (CT), and CD, each playing a vital role in maintaining the quality and reliability of software releases. The following sections will delve into the key aspects of traditional DevOps CI/CD pipelines, highlighting their components, tools, and best practices.

1) Continuous integration: CI involves the regular integration of code changes into a shared repository, a process that triggers automated builds and tests. This approach helps identify integration issues early, thereby reducing the time and effort needed to resolve them [7][8]. The market offers several common CI tools and technologies, including Jenkins, CircleCI, Travis CI, GitLab CI/CD, Bamboo, TeamCity, Azure DevOps, GitHub Actions, and Bitbucket Pipelines. Each tool provides unique features and integrations tailored to different aspects of software development workflows. The choice of a CI/CD tool often hinges on the specific needs and existing infrastructure of the development team. For instance, Jenkins offers unparalleled flexibility and plugin support, making it ideal for complex projects. CircleCI and Travis CI are known for their ease of use and quick setup, catering to cloud-based and GitHub-centric workflows, respectively. GitLab CI/CD and Azure DevOps provide comprehensive platforms for teams seeking integrated solutions, while GitHub Actions and Bitbucket Pipelines offer seamless experiences for users of their respective version control systems. These tools enhance the development process by automating builds, tests, and deployments, thereby improving collaboration and efficiency among team members. CI helps maintain a stable codebase, enhances code quality, and facilitates collaboration among development teams [8][22].

2) Continuous testing: CT is a vital practice in DevOps environments, ensuring software quality and reliability throughout the development life cycle. By integrating testing into the CI/CD pipeline, CT allows for early defect detection and accelerates feedback loops. Key strategies in continuous testing include Shift-Left Testing, Test-Driven Development (TDD), Continuous Test-Driven Development (CTDD), and Operational-Profile Based Testing. Shift-Left Testing moves testing activities to earlier stages in the development process, facilitating earlier defect identification and reducing the cost and effort required for corrections [9][10]. Test-Driven Development (TDD) and Continuous Test-Driven Development (CTDD) involve writing tests before the code itself, ensuring that development is guided by testing requirements. CTDD enhances this process by automating test execution and integrating it into the continuous testing framework [11]. Operational-Profile Based Testing uses data from software operations to guide testing, ensuring that tests reflect real-world usage patterns. This approach is particularly advantageous for reliability testing, as it helps evaluate software performance under actual operating conditions [12]. Automated testing tools are essential in the DevOps environment, enabling rapid and reliable software delivery through continuous integration and deployment processes. These tools support various testing types, including unit, integration, performance, and security testing, and are crucial for maintaining code quality and reliability. The choice of tools often depends on the project's specific needs, with factors to consider including platform compatibility, ease of use, and integration capability.

3) Continuous deployment: Continuous Deployment (CD) automates the release of validated code changes into production environments, ensuring the swift and reliable delivery of new features and fixes to users [8]. As a vital component of modern software development, CD enhances efficiency, reliability, and speed by automating application

deployment. A variety of tools and technologies support CD, each playing a distinct role within the deployment pipeline. These tools are crucial to the CI/CD process, facilitating seamless integration, testing, and deployment of software products. Git, a widely adopted version control system, tracks changes in source code during software development [13][14]. Jenkins, an open-source automation server, aids in building, deploying, and automating projects, often working alongside Git and Docker to streamline the CI/CD pipeline [14][15]. GitLab CI/CD, part of the GitLab platform, provides a robust CI/CD solution by integrating with Git repositories to enable automated testing and deployment [14]. Docker allows developers to package applications into containers, which are standardized units of software containing all necessary components to run an application, ensuring consistency across environments [15][13]. Kubernetes automates the deployment, scaling, and management of containerized applications, often collaborating with tools like ArgoCD to manage deployments within Kubernetes clusters [15][13]. While these tools are widely used in Continuous Deployment, the choice of tools may vary based on project requirements, team preferences, and organizational objectives. Integrating these tools into a unified CI/CD pipeline requires careful planning and execution to address potential challenges such as security vulnerabilities, system misconfigurations, and resource optimization. As software development evolves, adopting new tools and practices is essential for maintaining efficient and secure deployment.

4) Best practices and optimization: Optimizing CI/CD pipelines is essential for enhancing software development processes, as it boosts efficiency, reliability, and speed. As software systems become more complex, the demand for rapid and dependable deployments grows. Effective optimization of CI/CD pipelines results in notable improvements, such as increased deployment frequency, higher build success rates, and enhanced overall development efficiency. This process involves tackling several challenges, including inconsistencies in test environments, resource allocation issues, and build instabilities. By optimizing CI/CD pipelines, builds are stabilized, and execution efficiency is improved, which are critical factors for maintaining a seamless development workflow. This is achieved by addressing test environment inconsistencies and managing resources effectively [16]. Automated code integration and delivery minimize the time and errors associated with manual processes, enabling faster release cycles and allowing teams to focus more on business requirements [17]. CI/CD practices cultivate a culture of shared responsibility for code quality, enhancing collaboration among team members and boosting productivity [18]. Optimized pipelines ensure efficient scaling of software systems while maintaining reliability, which is crucial for handling complex software systems and large-scale projects [19] [20]. While optimization offers numerous benefits, it also presents challenges, such as managing the complexity of automation tools, ensuring toolchain compatibility, and addressing security concerns. Integrating emerging technologies like AI/ML into CI/CD processes can further enhance pipeline efficiency. Organizations must continuously adapt and refine their CI/CD strategies to remain competitive [21][20].

5) Summary: Traditional DevOps CI/CD pipelines form the foundation of modern software development practices, emphasizing automation, efficiency, and rapid delivery. These pipelines integrate core components of continuous integration, testing, and deployment, supported by various tools and technologies. This approach sets the stage for understanding how these principles are adapted in the context of machine learning operations.

B. Key Components of MLOps Pipelines

MLOps pipelines are essential for the effective deployment and management of machine learning models in production environments. They integrate various components to streamline the entire machine learning lifecycle, including data acquisition, model deployment, and monitoring. The primary components of MLOps pipelines are: data engineering, model development, CI, CT, CD, and governance. These components work in unison to ensure that machine learning models are scalable, reliable, and maintainable. While MLOps pipelines offer a structured approach to managing machine learning models, they encounter several challenges, such as talent shortages [25][41], interoperability issues [41], and regulatory compliance [23]. The success of MLOps hinges on its integration with business processes and adaptation to evolving industry standards [34]. Organizations must consider the interplay between technology, people, and processes to fully harness the potential of machine learning in driving business value [4].

1) Data engineering in MLOps: Data engineering is a vital aspect of MLOps, ensuring the smooth integration and operationalization of machine learning models. It systematically prepares and manages data, ensuring it is clean, consistent, and ready for model training. This includes automated processes for data cleaning, normalization, and transformation, which minimize manual intervention and errors [25]. Tools like Acumos and NiFi automate data pipelines, efficiently handling large datasets and ensuring continuous data updates for model training. Data engineering integrates into the machine learning pipeline, facilitating seamless data flow from ingestion to model deployment. This integration ensures the reproducibility and scalability of machine learning models. Modular pipelines and automated testing align data engineering processes with the overall MLOps framework, enabling continuous integration and delivery. Automation plays a significant role in MLOps data engineering. Tools like ALaaS implement automated workflows for data-centric AI tasks, reducing manual intervention and enhancing processing efficiency. Despite advancements, challenges persist in MLOps data engineering, including interoperability issues, regulatory compliance, and the need for continuous model training. Addressing these

challenges requires technical expertise and strategic planning [25]. Reusable MLOps frameworks, such as those offered by Acumos, provide solutions by enabling the reuse of existing infrastructure and deployment processes, thereby reducing the complexity and costs associated with data engineering tasks. Additional challenges include the need for continuous data updates and the integration of diverse data sources. The rapid evolution of machine learning technologies necessitates the ongoing adaptation of data engineering practices. However, organizations can overcome these challenges by leveraging automation and collaboration tools, achieving efficient and reliable data engineering processes.

2) Model development in MLOps: Within the MLOps framework, model development integrates DevOps principles to enhance machine learning processes. It promotes collaboration between data scientists and engineers, automates workflows, and ensures the continuous delivery of high-quality models [62][34]. CI/CD pipelines are vital in MLOps, enabling automated testing, validation, and deployment of models with minimal manual intervention [29][23]. Automation is a core aspect of MLOps, with tools and frameworks automating various stages of the model lifecycle, thereby reducing the time and effort needed to transition models from development to production [35]. Monitoring systems are essential for maintaining model performance in production environments, tracking predefined metrics to ensure models deliver accurate predictions and adapt to data changes [36].

3) Continuous integration in MLOps: CI is vital in MLOps, aiding the seamless integration of machine learning models into production environments. It automates testing and validation processes to ensure model quality prior to deployment. In MLOps, CI employs various tools, frameworks, and methodologies to optimize the machine learning lifecycle. Jenkins, an open-source CI tool, automates MLOps workflows by building pipelines for data analysis, preparation, training, testing, and deployment, thereby saving time and reducing manual effort for repetitive tasks [37]. Platforms like Kubeflow and MLflow offer end-to-end lifecycle management for ML applications, managing deployment pipelines and ensuring model version management and reproducibility [26][27]. ModelCI-e, a lightweight MLOps plugin, supports continuous integration and evolution by automating model updates and validation without requiring serving engine customization. It includes a model factory for prototyping and a backend for efficient orchestration of model updates [28]. CI pipelines in MLOps incorporate jobs to automatically train models and validate their performance, ensuring that only models meeting predefined quality standards are deployed, thus minimizing the risk of underperforming models [66]. Maintaining a centralized model registry and enforcing access controls are essential for managing model versioning and ensuring regulatory compliance, which are critical for scalable and robust MLOps pipelines [29]. Efficient resource management is imperative for CI in MLOps, as analyzing time and resource consumption in the ML pipeline helps identify potential performance bottlenecks, such as GPU (Graphics Processing Unit) utilization, which can impact CI process efficiency [27]. Addressing dynamic environments where online data diverges from offline training data presents a challenge, necessitating continuous learning and model updating techniques to maintain model relevance and performance [28]. While CI in MLOps offers benefits like enhanced model quality and reduced deployment time, it also presents challenges, including managing resource consumption and adapting to dynamic data environments. Addressing these challenges requires robust tools, efficient practices, and strategic planning for successful CI implementation.

4) Continuous testing in MLOps: CT is a crucial component of MLOps, ensuring the ongoing effectiveness and reliability of machine learning models in production. MLOps utilizes automated processes for retraining, deployment, and monitoring, enabling rapid iteration and adaptation to changing data and conditions [39][30][43][33]. MLOps pipelines, such as Continuous Training and Continuous Deployment-enabling (CTCD-e), automate model retraining and redeployment, triggering retraining when performance declines and conducting A/B testing to ensure optimal model functionality. CI and CD practices are tailored for ML workflows to facilitate efficient model updates, automating the entire lifecycle from data ingestion to deployment [29]. Systems like ModelCI-e support continual learning by automating model updates and validation [28]. Comprehensive monitoring mechanisms are essential for tracking model performance and detecting data drift, ensuring models remain accurate and reliable over time [29]. MLOps processes must incorporate operational feedback to continuously innovate and adapt models. While continuous testing in MLOps offers significant benefits, it also presents challenges. Automating various stages of the MLOps process requires robust infrastructure and tools. Ensuring model reproducibility and traceability is vital for maintaining trust and accountability [24]. Additionally, integrating MLOps with existing IT and operational systems can be complex. In conclusion, continuous testing in MLOps combines automated pipelines, continual learning, and robust monitoring systems. These practices ensure machine learning models remain effective in dynamic production environments. Implementing continuous testing requires addressing challenges related to automation, safety assurance, and system integration. Overcoming these challenges allows organizations to fully leverage MLOps benefits, enhancing the scalability, reliability, and productivity of their machine learning systems.

5) Continuous deployment in MLOps: CD within MLOps is vital for the efficient and reliable deployment of machine learning models into production environments. It automates the deployment pipeline, facilitating seamless updates and model integration, thereby enhancing operational efficiency and reducing time-to-market. Integrating CI and CD pipelines

is crucial for maintaining model performance in dynamic environments. Automation is a cornerstone of CD in MLOps, streamlining the entire lifecycle of ML models, from data ingestion to monitoring. Tools such as MLflow, Kubeflow, and Airflow manage deployment pipelines, ensuring consistent and efficient model deployment. Automated cycles convert code changes into container images, which are then deployed to production environments [31]. Comprehensive monitoring and observability mechanisms track model performance and detect drift, maintaining the trustworthiness of ML models in production. Predefined metrics ensure prediction quality throughout deployment, facilitating continuous improvement [32]. Continuous deployment in MLOps offers benefits like increased efficiency and reduced time-to-market, but it also presents challenges [41][23][29][26][4]. These challenges include managing dependency complexity and ensuring model reproducibility and traceability. Best practices for addressing these challenges involve maintaining a centralized model registry and enforcing access controls. Ensuring compliance with regulatory requirements is also crucial. Collaboration among data scientists, engineers, and business stakeholders fosters innovation and agility in model deployment.

6) Governance in MLOps: Governance is a cornerstone of MLOps, ensuring the responsible, ethical, and compliant deployment of machine learning models. It involves practices and policies that manage the lifecycle of ML models in accordance with organizational standards and regulatory requirements. Governance in MLOps is vital for maintaining model integrity, protecting data privacy, and building trust in AI systems. This section explores the key aspects of governance in MLOps: compliance, ethical considerations, and model management. Compliance in MLOps entails adhering to legal and regulatory frameworks, such as data protection laws and industry-specific standards. Organizations must establish access controls and audit trails to monitor model access and modifications, ensuring accountability and traceability [29][23]. A centralized model registry is crucial for tracking model versions and changes, facilitating audits and regulatory reporting [29]. Ethical considerations are a vital component of governance, emphasizing fairness, transparency, and accountability in machine learning models [25][38]. Governance frameworks should include guidelines for detecting and mitigating bias to prevent the perpetuation or exacerbation [38]. Transparency in model decision-making is essential, and organizations should implement explainable AI techniques to make model outputs understandable to stakeholders [38]. Effective governance requires comprehensive model management practices, including version control, performance monitoring, and drift detection [29][39]. Continuous monitoring of model performance is crucial for identifying issues like data drift or model degradation, enabling timely interventions [29]. Governance frameworks should also include policies for model retraining and updates to maintain relevance and effectiveness over time [39]. A significant challenge in MLOps governance is balancing innovation and compliance. Overly stringent governance can stifle creativity and delay model deployment [25]. Proposed solutions include implementing flexible governance frameworks and using automation to streamline processes [29][4]. Effective collaboration among data scientists, engineers, and compliance officers is crucial for developing practical and effective governance policies [23]. While essential, governance in MLOps faces additional challenges. The dynamic nature of AI technologies and the evolving regulatory landscape complicate compliance efforts. There is also a need for more standardized governance frameworks adaptable across various industries and use cases. Despite these challenges, effective governance remains fundamental to successful MLOps practices, enabling organizations to leverage AI responsibly and ethically.

7) Summary: MLOps CI/CD pipelines build upon traditional DevOps principles, tailoring them to the unique requirements of machine learning projects. These pipelines encompass data engineering, model development, CI/CD, and governance, addressing specific challenges in ML model production. The complexities and ethical considerations in MLOps highlight the need for specialized approaches in managing ML systems throughout their lifecycle.

IV. Challenges in Implementing Machine Learning Components Within Larger Applications

The integration of ML components into larger applications is becoming increasingly common, driven by the potential to enhance decision-making, automate processes, and deliver personalized experiences across various fields. However, this integration is not without its challenges. Technical complexities present significant hurdles, while organizational and cultural barriers further impede successful implementation. Addressing these issues necessitates a holistic approach to ensure the effective deployment and maintenance of ML components. This response examines the key challenges in integrating ML components and offers actionable strategies to overcome these obstacles. The insights presented are drawn from relevant literature in the field.

A. Technical Challenges in ML Integration

1) Model deployment and monitoring: The deployment and monitoring of ML models in production environments present significant technical challenges. Many organizations find it difficult to design architectures for production deployment and to integrate ML models into legacy systems [40][41]. Inadequate monitoring practices often result in poor tracking of models in production, leading to performance degradation over time [40][42]. To address these issues, [43] and [41] recommend implementing robust MLOps practices, which include: using version control systems for model versioning, employing containerization for consistent environments, and continuously monitoring model performance. Tools like Kubeflow can automate several processes, such as hyperparameter tuning, model deployment,

and maintenance. These automated processes reduce manual effort and enhance reliability [44].

2) Cross-Platform integration: The authors in [45] underscore a major technical hurdle: the integration of ML models across a range of platforms and technologies. These models are typically developed within specialized environments, and deploying them on new platforms such as Metal, Vulkan, or WebGPU demands considerable effort and customization. The authors suggest a solution through a top-down development approach, as demonstrated by TapML. This method streamlines the deployment of ML models across various platforms by utilizing automated testing and progressively transferring computations to the target platforms. This strategy effectively minimizes the need for extensive debugging and validation efforts.

3) Data quality and versioning: Ensuring data quality and versioning is essential for the success of ML models [46][47]. Inadequate data quality can result in models that are biased or inaccurate. Additionally, concept drift, which refers to changes in data distributions, can render models ineffective over time. To tackle these issues, experts advocate for thorough data preprocessing and versioning [41][47]. Incorporating data version control into the ML lifecycle is crucial for maintaining model effectiveness. Automated data validation pipelines are also vital in upholding data quality. These methods enable teams to monitor changes over time and ensure high-quality data inputs. By adopting these practices, organizations can enhance the reliability and longevity of their ML models.

4) Maintainability, scalability, and reliability: Microservices architecture enables the development of ML components as standalone services, which can be seamlessly integrated with other system components through well-defined application programming interfaces (APIs). This modular approach allows for the smooth incorporation of ML models into larger systems, facilitating updates or replacements without affecting other components [48][49]. By employing common communication protocols like Representational State Transfer (REST) over Hypertext Transfer Protocol (HTTP), microservices ensure interoperability among diverse components, even those developed in different programming languages or frameworks. This is particularly crucial for integrating ML models, which may require specific environments or dependencies [50][51]. Each microservice, including those for ML, can be developed, tested, and deployed independently, reducing the complexity of managing large codebases and allowing for more frequent updates and bug fixes, thereby enhancing maintainability [52][53]. Encapsulating ML models as microservices promotes code reusability and simplifies maintenance. Changes to a model or its underlying algorithms can be made without impacting other services, streamlining the maintenance process [48]. Microservices architecture supports the independent scaling of services based on demand. ML services, which often require significant computational resources, can be scaled independently of other system components, optimizing resource utilization and ensuring efficient handling of large data volumes [52][50]. The cloud-native nature of microservices allows for automated scaling and elasticity, essential for managing the variable workloads typical of ML applications. Moreover, the isolation of services ensures that failures in one service do not affect others, thereby enhancing system reliability [50][54].

While microservices architectures offer substantial advantages, they also present challenges, such as increased communication overhead and potential network delays due to the distributed nature of services. Ensuring data consistency and managing the complexity of service orchestration are critical considerations. Additionally, integrating ML models requires careful design to address issues related to model versioning and deployment pipelines [52][55]. Despite these challenges, the benefits of microservices in terms of scalability and maintainability make them a compelling choice for integrating ML components into complex systems.

B. Organizational Challenges in ML Integration

1) Collaboration between cross-functional teams: The authors in [56] contend that machine learning-powered systems require collaboration among data scientists, software engineers, and domain experts. However, effective teamwork often encounters challenges, such as disparities in technical expertise, ambiguous roles, and insufficient communication. To address these issues, the authors propose several solutions. First, they recommend establishing clear roles and responsibilities. Second, they suggest fostering a collaborative culture. Third, they advise utilizing communication tools to enhance teamwork. Additionally, the authors in [56] highlight the importance of concise system documentation, which can help bridge the gap between data scientists and software engineers, facilitating better understanding and cooperation.

2) Managing diverse tools and frameworks: The machine learning ecosystem encompasses a wide array of tools and frameworks [41][47]. This diversity can pose challenges in terms of integration and maintenance. Organizations often encounter significant obstacles in managing these tools while ensuring consistency across different environments. To tackle these issues, the authors in [41] and [57] advocate for the standardization of tools and frameworks whenever feasible. They also recommend employing platform-independent execution frameworks to minimize complexity. Furthermore, the implementation of automated pipelines for model training and deployment can help streamline workflows.

3) Integrating ML workflows with existing processes: Integrating ML workflows into existing software development processes poses considerable challenges, especially for organizations with legacy systems [43][41]. This integration is particularly complex when dealing with continuous CI/CD pipelines. To address these challenges, adopting MLOps practices can facilitate a more seamless integration. These practices align with DevOps principles and offer practical solutions. One key approach is the utilization of versioned

environments. Another effective strategy is the implementation of containerization. Both methods ensure consistency and reproducibility in ML workflows.

C. Cultural Challenges in ML Integration

1) Resistance to change: Resistance to adopting ML technologies is a prevalent cultural challenge. Many stakeholders remain skeptical about ML's value, while others hesitate to replace traditional methods with data-driven approaches [58]. Overcoming this resistance necessitates targeted educational efforts. Stakeholders must understand ML's benefits and recognize its practical value. Pilot projects can effectively demonstrate this, offering tangible evidence of ML's potential. Additionally, fostering a culture that encourages continuous learning and experimentation can support broader acceptance [41][58].

2) Skill gaps: Effectively integrating ML components often demands specialized expertise. Conventional software development teams may lack these skills, which can impede adoption and restrict collaboration between data scientists and engineers [41][47]. To bridge this gap, organizations should invest in upskilling initiatives. Promoting collaboration between data scientists and software engineers can also enhance integration efforts. Furthermore, cross-functional training and knowledge-sharing programs contribute to building more cohesive and capable teams [41][56].

3) Lack of shared understanding: A thorough understanding of ML concepts and their practical application is crucial for their successful integration into educational curricula. However, stakeholders often possess varying levels of comprehension, which can result in misaligned expectations [59]. Perspective-based approaches, such as PerSpecML, are instrumental in aligning these expectations. They ensure that all stakeholders have a clear understanding of the system's goals, user experience, and technical requirements [59]. Integrating ML components into broader applications presents complex challenges that span technical, organizational, and cultural domains. Overcoming these challenges requires more than just technical expertise; it also necessitates effective organizational strategies and intentional cultural adaptation. Organizations can tackle these issues by adopting MLOps frameworks, fostering cross-functional collaboration, and cultivating a shared understanding of ML principles. These efforts enable teams to navigate integration challenges and fully harness the benefits of machine learning. A summary of the challenges and corresponding solutions is provided in Table I.

TABLE I. KEY CHALLENGES AND SOLUTIONS IN ML INTEGRATION

Challenge	Description	Solution
Model Deployment	Difficulty in deploying models in production environments and integrating with legacy systems.	Implement MLOps practices, including version control, containerization, and continuous monitoring [43][40][41].
Cross-Platform Integration	Challenges in deploying models on diverse platforms like Metal, Vulkan, or WebGPU.	Use top-down development approaches like TapML for streamlined deployment [45].
Data Quality and Versioning	Poor data quality and concept drift leading to model degradation.	Implement data versioning and automated validation pipelines [41][47].
Maintainability, Scalability, and Reliability	Efficient updates, handling of increasing data and workload demands, and consistent delivery of accurate and dependable results.	Using a microservices architecture enables updates or replacements to be made without affecting other components [48][49].
Collaboration	Differences in expertise and unclear roles hindering teamwork.	Foster collaboration through clear roles, communication tools, and system documentation [56].
Tool Management	Managing diverse ML tools and frameworks.	Standardize tools and adopt automated pipelines for consistency [41][57].
Cultural Resistance	Stakeholder skepticism and resistance to ML adoption.	Educate stakeholders and build a culture of continuous learning [41][58].
Skill Gaps	Lack of specialized skills for ML implementation.	Invest in upskilling programs and cross-functional training [41][56].

V. EMERGING FRAMEWORKS AND METHODOLOGIES

The integration of ML into applications is facilitated by several emerging frameworks and methodologies, each offering unique advantages and challenges. These frameworks and methodologies streamline the ML lifecycle, enhance collaboration, and ensure the reliability and scalability of ML models. Table II presents the suitability of various approaches for ML integration within applications:

TABLE II. ML INTEGRATION FRAMEWORKS

Framework/Methodology	Description
CRISP-DM (Cross-Industry Standard Process for Data Mining) [24].	- Offers a structured approach to ML projects. - Focuses on business understanding, data preparation, modeling, evaluation, and deployment. - May lack agility and continuous integration features for modern ML applications. - Less suitable for dynamic environments requiring rapid iterations
DataOps [61].	- An agile methodology aimed at improving data analytics quality and reducing cycle time. - Emphasizes collaboration, automation, and monitoring. - Crucial for managing data pipelines in ML applications. - Can be integrated with MLOps to enhance data management and operational efficiency.
MLOps Frameworks [60] [62][63].	- Integrate ML, DevOps, and data engineering to automate and enhance the ML lifecycle. - Facilitate continuous integration, delivery, and monitoring. - Ensure model reliability and scalability. - Examples include Kubernetes-based open-source frameworks and proprietary solutions like Amazon SageMaker.

Framework/Methodology	Description
Agile ML [24]	- Applies agile principles to ML development. - Promotes iterative development and collaboration. - Advantageous in environments requiring rapid prototyping and frequent updates. - Enables teams to swiftly adapt to changes - May face challenges in maintaining model stability due to frequent changes.
Feature Stores [64][65]	- Centralized repositories for storing and managing ML features. - Facilitate feature reuse, consistency, and governance. - Crucial for scalable ML applications. - Enhance collaboration between data scientists and engineers.
Site Reliability Engineering (SRE) for ML Systems [64]	- Adapted to ensure the reliability and performance of ML systems. - Emphasizes monitoring, incident response, and performance optimization. - Critical for production-grade ML applications. - Can be integrated with MLOps to enhance robustness and trustworthiness.
Continuous Training (CT) ([24][66]	- Involves perpetual retraining of ML models to accommodate new data and evolving environments. - Crucial for applications with rapidly changing data. - Ensures models maintain accuracy and relevance. - Requires robust data pipelines and monitoring systems to manage model drift and performance degradation.

These frameworks and methodologies offer significant benefits for ML integration but also present challenges. For instance, integrating Agile ML and Continuous Training requires careful management to prevent model instability. The choice of framework or methodology should align with the specific needs and constraints of the application. Factors to consider include the need for rapid iteration and the importance of model reliability. Successful integration of ML into applications depends on selecting the right combination of frameworks and methodologies that best fit the project's goals and requirements.

VI. Security and Governance in ML Systems

Security and governance are crucial for building trust in ML systems. Protecting against adversarial attacks and data breaches is essential for reliability and acceptance, particularly in sensitive sectors like healthcare and finance [67][68]. ML systems are susceptible to adversarial attacks, where malicious inputs can deceive the model, compromising its integrity and reliability [69]. Safeguarding data privacy is vital, as unauthorized access and breaches can lead to the misuse of sensitive information [70]. The ethical use and accountability of ML systems are significant concerns, with governance frameworks playing a key role in ensuring ethical application and establishing accountability. This involves developing policies and standards for ethical deployment [71][72]. Additional concerns include model theft and intellectual property protection, as attackers may attempt to extract model parameters or replicate functionality, threatening intellectual property [69]. Developing comprehensive governance frameworks is challenging due to rapid technological advancements [71]. Integrating security into the ML lifecycle through Secure Machine Learning Operations (SecMLOps) can enhance system security and reliability by incorporating security measures from the design phase throughout the system's lifecycle [73]. Advanced security techniques, such as adversarial training, model hardening, and secure computing environments, can mitigate risks in ML workflows (Chittibala & Jabbireddy, 2024). Robust governance frameworks should promote transparency, accountability, and compliance with legal and ethical standards [71][72]. Effective management of the ML system lifecycle, including regular updates and patches, is crucial for maintaining security and functionality over time [72]. While security and governance are vital, potential trade-offs and challenges may arise. Stringent security measures could impact system performance and usability, and rapid technological advancement may outpace the development of governance frameworks, leading to regulatory gaps. Continuous research and adaptation of security and governance practices are essential to keep pace with the evolving ML landscape.

VII. Limitations of the Study

As a critical review, this study synthesizes existing literature rather than generating new empirical data. While this approach is effective for providing a comprehensive overview and identifying gaps, it relies on the quality and scope of the published research available at the time of writing. The findings are thus reflective of the current state of the art as documented in the literature. The review focuses specifically on the integration of ML into software engineering practices, with an emphasis on the evolution from DevOps to MLOps, challenges, and emerging solutions. While an effort was made to be comprehensive, the rapid evolution of ML technologies and software engineering practices means that some emerging trends or niche applications might not be fully captured. The generalizability of specific solutions may vary depending on organizational context, industry, and scale of ML implementation. Although a systematic approach was intended, the selection of literature for review might inherently carry some bias. The emphasis on certain frameworks, tools, or challenges could be influenced by their prominence in the academic and industry discourse, potentially underrepresenting less-documented but equally valid perspectives or solutions. The field of ML integration in software engineering is characterized by continuous innovation and rapid advancements. While the study aims to provide an up-to-date overview, new tools, practices, or challenges may emerge quickly, potentially altering the landscape described. This inherent dynamism means that any review, by its nature, offers a snapshot of a constantly moving target.

VIII. Conclusion

The integration of ML into software engineering practices has ushered in a new era of development and operational processes, fundamentally transforming how organizations approach their technological strategies. This shift from traditional DevOps to MLOps frameworks represents a significant leap forward, addressing the unique challenges posed by deploying and maintaining ML models in production

environments. As organizations increasingly leverage ML to gain competitive advantages, the importance of robust MLOps pipelines cannot be overstated.

The journey towards effective ML integration is marked by both opportunities and challenges. On one hand, ML offers unprecedented capabilities for data analysis, prediction, and automation. On the other hand, it introduces complexities in data management, model versioning, and continuous monitoring that demand sophisticated solutions. The emergence of frameworks like CRISP-DM, DataOps, and Agile ML provides potential pathways for organizations to navigate these challenges, offering structured approaches to streamline the ML lifecycle.

Looking forward, the field of ML integration in software engineering is poised for rapid evolution. As organizations grapple with the intricacies of deploying large language models (LLMs) and other advanced ML systems, new strategies are emerging to address computational demands, integration complexities, and ethical concerns. The development of more sophisticated MLOps frameworks, coupled with advancements in model interpretability and security measures, will be crucial in shaping the future landscape of ML-driven software engineering.

The success of ML integration will ultimately hinge on an organization's ability to adopt a holistic approach that encompasses technical, organizational, and cultural dimensions. This includes fostering cross-team collaboration, implementing robust governance frameworks, and maintaining a commitment to ethical AI practices. As the field continues to advance, organizations must remain adaptable, investing in continuous learning and skill development to stay ahead of the curve.

In essence, the integration of ML into software engineering practices represents both a challenge and an opportunity for innovation. By embracing flexible, integrated frameworks and maintaining a focus on security, governance, and ethical considerations, organizations can harness the full potential of ML to drive technological progress and business success in an increasingly data-driven world.

REFERENCES

[1] Rajak, S. K., Kumari, S., Kumar, M., & Siddharth, D., "Application of machine learning for software engineers," In A. Sharma, N. Chanderwal, A. Prajapati, P. Singh, & M. Kansal (Eds.), Advancing Software Engineering Through AI, Federated Learning, and Large Language Models. IGI Global Scientific Publishing. pp. 54-69, 2024, https://doi.org/10.4018/979-8-3693-3502-4.ch004.

[2] Zhang, D., Tsai, J.J., "Machine learning and software engineering," Software Quality Journal, vol. 11, pp. 87–119, June 2003, https://doi.org/10.1023/A:1023760326768.

[3] Wang, C., Chen, Z., & Zhou, M., "AutoML from software engineering perspective: landscapes and challenges," 2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR), Melbourne, Australia, pp. 39-51, July 2023, https://doi.org/10.1109/MSR59073.2023.00019.

[4] Eken, B., Pallewatta, S., Tran, N. K., Tosun, A., & Babar, A., A "Multivocal review of MLOps practices, challenges, and open issues," ACM Computing Surveys, vol 58, iss 2, pp 1-35, September 2025, https://dl.acm.org/doi/10.1145/3747346.

[5] Wang, S., Huang, L., Gao, A., Ge, J., Zhang, T., Feng, H., Satyarth, I., Li, M., Zhang, H., Ng, V., "Machine/deep learning for software engineering: a systematic literature review." in IEEE Transactions on Software Engineering, vol. 49, no. 3, pp. 1188-1231, March 2023, https://doi.org/10.1109/TSE.2022.3173346.

[6] Kotti, Z., Galanopoulou, R., & Spinellis, D., "Machine learning for software engineering: a tertiary study," ACM Computing Surveys, vol. 55, iss. 12, March 2023, https://doi.org/10.1145/3572905.

[7] Paul, A., Haldar, M., "Continuous integration and continuous delivery/continuous deployment," in Serverless Web Applications with AWS Amplify. Apress, Berkeley, CA. pp. 223-256, August 2023, https://doi.org/10.1007/978-1-4842-8707-1_8.

[8] Istifarulah, M. H. R., & Tiaharyadini, R., "DevOps, continuous integration, and continuous deployment methods for software deployment automation," JISA (Jurnal Informatika Dan Sains), vol. 6, iss. 2, December 2023, https://doi.org/10.31326/jisa.v6i2.1751.

[9] Shinde, B. B., "Automated testing in DevOps: strategies and tools," International Journal of Advanced Research in Science, Communication and Technology, vol. 4, iss. 4, June 2024, https://doi.org/10.48175/ijarsct-19074.

[10] Angara, J., Gutta, S., & Prasad, S., "DevOps with continuous testing architecture and its metrics model," in Sa, P., Bakshi, S., Hatzilygeroudis, I., Sahoo, M. (eds). Recent Findings in Intelligent Computing Techniques. Advances in Intelligent Systems and Computing Springer, Singapore, vol 709, pp. 271-281, November 2018, https://doi.org/10.1007/978-981-10-8633-5_28.

[11] Madeyski, L. and Kawalerowicz, M., "Continuous Test-Driven Development - A Novel Agile Software Development Practice and Supporting Tool," in Proceedings of the 8th International Conference on Evaluation of Novel Approaches to Software Engineering - ENASE; ISBN 978-989-8565-62-4; ISSN 2184-4895, SciTePress, pp. 260-267, July 2013, https://doi.org/10.5220/0004587202600267.

[12] Pietrantuono, R., Bertolino, A., Angelis, G. D., Miranda, B., & Russo, S., "Towards continuous software reliability testing in DevOps," 2019 IEEE/ACM 14th International Workshop on Automation of Software Test (AST), Montreal, QC, Canada, pp. 21-27, September 2019, https://doi.org/10.1109/AST.2019.00009.

[13] Singh Nikhil, Patel Durgesh, Raj Ankit, Shubham, Kour Sukhmeet, "CI/CD pipeline for web applications," International Journal For Research in Applied Science and Engineering Technology, vol. 11, iss. V, May 2023 https://doi.org/10.22214/ijraset.2023.52867.

[14] Babenko, V., Taraniuk, V., Tkachenko, V., & Klymenko, I., "CI/CD integration tools for automated code deployment and verification for training purposes," Information, Computing and Intelligent Systems, iss. 5, December 2024. https://doi.org/10.20535/2786-8729.5.2024.318795.

[15] Shrestha, R., & Ray, A. K., "Streamlining application deployment: a CI/CD pipeline for Kubernetes," 2024 IEEE International Conference on Cloud Engineering (IC2E), Paphos, Cyprus, pp. 253-255, September 2024, https://doi.org/10.1109/ic2e61754.2024.00038.

[16] Shriram, K. M. P., "Engineering efficiency through CI/CD pipeline optimization," International Journal of Science and Research Archive, vol. 14, iss. 1, pp. 908-916, January 2025, https://doi.org/10.30574/ijsra.2025.14.1.0107.

[17] Namsraidorj, M., Lkhaasuren, S., Gendensuren, B., Radnaa, K., Rentsendorj, J., & Enkhtur, A., "Continuous integration and delivery of software products: pipeline implementation," International Journal of Engineering and Computer Science, vol. 13, no. 5. 2024, https://doi.org/10.18535/ijecs/v13i05.4821.

[18] Gujar, S., & Patil, S., "Continuous integration and continuous deployment (CI/CD) optimization," International Journal of Innovative Science and Research Technology, vol. 9, iss. 10, October 2024, https://doi.org/10.38124/ijisrt/ijisrt24oct014.

[19] Emmanni, P.S., "Implementing CI / CD pipelines for enhanced efficiency in IT projects," International Journal of Science and Research, vol. 9, iss. 9, September 2020, https://doi.org/10.21275/sr24402001528.

[20] Soma, V., "Enhancing CI/CD pipelines with Azure pipelines," Journal of Engineering and Applied Sciences Technology, vol. 6, iss. 8, August 2024, https://doi.org/10.47363/jeast/2024(6)e108.

[21] Thatikonda, V. K., "Beyond the buzz: a journey through CI/CD principles and best practices," European Journal of Theoretical and

Applied Sciences, vol. 1, no. 5, 2023. https://doi.org/10.59324/ejtas.2023.1(5).24.

[22] Dileepkumar, S. R., & Mathew, J. "Enhancing DevOps and continuous integration in software engineering: a comprehensive approach," the 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichirappalli, India, pp. 01-05, June 2023, https://doi.org/10.1109/ICEEICT56924.2023.10157286.

[23] Immaneni, J., "Building MLOps pipelines in fintech: keeping up with continuous machine learning," International Journal of Science and Research (IJSR), vol 9, iss. 3, pp. 1726-1734, March 2020, https://www.doi.org/10.21275/sr20034093248.

[24] Testi, M., Ballabio, M., Frontoni, E., Iannello, G., Moccia, S., Soda, P., and Vessio, G., "MLOps: a taxonomy and a methodology," in IEEE Access, vol. 10, pp. 63606-63618, June 2022, https://doi.org/10.1109/ACCESS.2022.3181730.

[25] Sachdeva, M. S., "MLOps: Revolutionizing AI Development and Deployment," International Journal For Multidisciplinary Research, vol. 6, iss. 5. September 2024, https://doi.org/10.36948/ijfmr.2024.v06i05.28794.

[26] Jana, A. K., Saha, S., "The MLOps approach to model deployment: a road map to seamless scalability," Journal of Artificial Intelligence & Cloud Computing, vol. 1, iss. 1, pp. 1-4, February 2022, https://doi.org/10.47363/jaicc/2022(1)267.

[27] Zhou, Y., Yu, Y., & Ding, B., "Towards MLOps: a case study of ML pipeline platform," 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE), Beijing, China, pp. 494-500, March 2021, https://doi.org/10.1109/ICAICE51518.2020.00102.

[28] Huang, Y., Zhang, H., Wen, Y., Sun, P., & Ta, N. B. D., "ModelCI-e: enabling continual learning in deep learning serving systems," Cornell University, Distributed, Parallel, and Cluster Computing, June 2021, https://arxiv.org/abs/2106.03122.

[29] Vijayan, N. E., "Building scalable MLOps: optimizing machine learning deployment and operations," International Journal of Scientific Research In Engineering and Management (IJSREM), vol. 8, iss. 10, October 2024, https://doi.org/10.55041/ijsrem37784.

[30] Luo, Y., Raatikainen, M., and Nurminen, J. K., "Autonomously adaptive machine learning systems: experimentation-driven open-source pipeline," 2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), Durres, Albania, pp. 44-52, January 2024, https://doi.org/10.1109/SEAA60479.2023.00016.

[31] Chowdary, M. N., Sankeerth, B., Chowdary, C. K., & Gupta, M., "Accelerating the machine learning model deployment using MLOps'. 4th International Conference on Intelligent Circuits and Systems. Journal of Physics: Conference Series, vol. 2327, April 2022 https://doi.org/10.1088/1742-6596/2327/1/012027.

[32] Bodor, A., Hnida, M., & Daoudi, N., "From development to deployment: an approach to MLOps monitoring for machine learning model operationalization," 2023 14th International Conference on Intelligent Systems: Theories and Applications (SITA), Casablanca, Morocco, pp. 1-7, January 2024, https://doi.org/10.1109/SITA60746.2023.10373733.

[33] Bodor, A., Hnida, M., & Daoudi, N., "Machine learning models monitoring in MLOps context: metrics and tools," International Journal of Interactive Mobile Technologies (iJIM), vol.17, iss. 23, pp. 125–139, December 2023, https://doi.org/10.3991/ijim.v17i23.43479.

[34] Haertel, C., Staegemann, D., Daase, C., Pohl, M., Nahhas, A., & Turowski, K., "MLOps in data science projects: a review," 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, pp. 2396-2404, January 2024, https://doi.org/10.1109/BigData59044.2023.10386139.

[35] Mahida, A., "A review on continuous integration and continuous deployment (CI/CD) for machine learning," International Journal of Science and Research (IJSR), vol. 10, no. 3, pp. 1967-1970, 2021, https://doi.org/10.21275/sr24314131827.

[36] Elgamal, Z. S., Elfangary, L., & Fahmy, H. (2024), "A machine learning operations (MLOps) monitoring model using BI-LSTM and SARSA algorithms, " International Journal of Advanced Computer Science and Applications, vol. 15, iss. 10, 2024, https://doi.org/10.14569/ijacsa.2024.0151060.

[37] R, N., & Mohana, M., "Jenkins pipelines: a novel approach to machine learning operations (MLOps)," 2022 International Conference on Edge Computing and Applications (ICECAA), Tamilnadu, India, pp. 1292-1297. November 2022, https://doi.org/10.1109/ICECAA55415.2022.9936252.

[38] Kodakandla, N., "Scaling AI responsibly: leveraging MLOps for sustainable machine learning deployments," International Journal of Science and Research Archive, vol. 13, iss. 1, pp. 3447-3455, October 2024, https://doi.org/10.30574/ijsra.2024.13.1.1798.

[39] Yemane, M., "MLOps for PHM systems," Proceedings of the Asia Pacific Conference of the PHM Society 2023, vol. 4 no. 1, September 2023, https://doi.org/10.36001/phmap.2023.v4i1.3703.

[40] Zimelewicz, E., Kalinowski, M., Méndez, D., Giray, G., Alves, A. P. S., Lavesson, N., Azevedo, K., Villamizar, H., Escovedo, T., Lopes, H., Biffl, S., Musil, J., Felderer, M., Wagner, S., Baldassarre, M. T., & Gorschek, T., "ML-Enabled systems model deployment and monitoring: status quo and problems," in: Bludau, P., Ramler, R., Winkler, D., Bergsmann, J. (eds) Software Quality as a Foundation for Security. SWQD 2024. Lecture Notes in Business Information Processing, Springer, Cham., vol 505, pp. 112-131, Apil 2024, https://doi.org/10.1007/978-3-031-56281-5_7.

[41] Singla, A., "Machine learning operations (MLOps): challenges and strategies," Journal of Knowledge Learning and Science Technology, vol 2, iss 3, 2023, Online. https://doi.org/10.60087/jklst.vol2.n3.p340.

[42] Shankar, S., Garcia, R., Hellerstein, J. M., & Parameswaran, A. G., ""We have no idea how models will behave in production until production": how engineers operationalize machine learning," Proceedings of the ACM on Human-Computer Interaction, vol 8, iss CSCW1, article 206, pp. 1-34, April 2024, https://doi.org/10.1145/3653697.

[43] Liang, P., Song, B., Zhan, X., Zhou, C., & Yuan, J., "Automating the training and deployment of models in MLOps by integrating systems with machine learning," Proceedings of the 2nd International Conference on Software Engineering and Machine Learning, 2024, https://doi.org/10.54254/2755-2721/76/20240690.

[44] Hsu, C.-C., Chen, P., & Wu, I.-Z, "End-to-End automation of ML model lifecycle management using machine learning operations platforms," 2024 International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan), Taichung, Taiwan, 2024 pp. 209-210, September 2024, doi: https://doi.org/10.1109/ICCE-Taiwan62264.2024.10674445.

[45] Feng, S., Liu, J., Lai, R., Ruan, C. F., Yu, Y., Zhang, L., & Chen, T., "Emerging Platforms Meet Emerging LLMs: A Year-Long Journey of Top-Down Development," April 2024, https://arxiv.org/pdf/2404.09151v2 or https://arxiv.org/html/2404.09151v2.

[46] Nahar N., Zhang H., Lewis G., Zhou S. and Kästner C, "A Meta-Summary of Challenges in Building Products with ML Components – Collecting Experiences from 4758+ Practitioners," 2023 IEEE/ACM 2nd International Conference on AI Engineering – Software Engineering for AI (CAIN), Melbourne, Australia, pp. 171-183, July 2023, https://doi.org/10.1109/cain58948.2023.00034.

[47] Rahman, M.S., Khomh, F., Hamidi, A., Cheng, J., Antoniol, G., Washizaki, H., "Machine learning application development: practitioners' insights," Software Quality Journal, vol. 31, 1065–1119, March 2023, https://doi.org/10.1007/s11219-023-09621-9.

[48] Ribeiro, J. L., Figueredo, M., Araujo, A. de, Cacho, N., & Lopes, F., "A Microservice based architecture topology for machine learning deployment," 2019 IEEE International Smart Cities Conference (ISC2), Casablanca, Morocco, pp. 426-431, April 2020, https://doi.org/10.1109/ISC246665.2019.9071708.

[49] He, Y., Zhang, Y., Wu, C., Yang, M., Xu, W., Wan Haiyang, "Architecture design and application of IIoT platform in automobile manufacturing based on microservices and deep learning techniques," in IEEE Access, vol. 12, pp. 166834-166842, October 2024, https://doi.org/10.1109/ACCESS.2024.3487832.

[50] Hasselbring, W., "Microservices for scalability: keynote talk abstract," ICPE '16: Proceedings of the 7th ACM/SPEC on International

Conference on Performance Engineering, pp. 133–134, March 2016, https://doi.org/10.1145/2851553.2858659.

[51] Valdivia, J. A., Lora-González, A., Limón, X., Cortes-Verdin, K., & Ocharán-Hernández, J. O., "Patterns related to microservice architecture: a multivocal literature review," Programming and Computer Software, vol. 46, pp. 594-608, December 2020, https://doi.org/10.1134/S0361768820080253.

[52] Thatikonda, V., "Assessing the impact of microservices architecture on software maintainability and scalability," European Journal of Theoretical and Applied Sciences, vol. 1, no. 4, 2023, https://doi.org/10.59324/ejtas.2023.1(4).71.

[53] Hasselbring, W., & Steinacker, G., "Microservice architectures for scalability, agility, and reliability in e-commerce," 2017 IEEE International Conference on Software Architecture Workshops (ICSAW), Gothenburg, Sweden, pp. 243-246, June 2017, https://doi.org/10.1109/ICSAW.2017.11.

[54] Vaidhyanathan, K., Caporuscio, M., Florio, S., & Muccini, H., "ML-enabled service discovery for microservice architecture: a QoS approach," SAC'24: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, pp. 1193 – 1200, May 2024, https://doi.org/10.1145/3605098.3635942.

[55] Weerasinghe, L., & Perera, I. (2024).Reference Architecture for Microservices with an Optimized Inter-Service Communication Strategy. 2024 International Research Conference on Smart Computing and Systems Engineering (SCSE), Colombo, Sri Lanka, 2024, pp. 1-6, https://doi.org/10.1109/scse61872.2024.10550466.

[56] Busquim, G., Villamizar, H., Lima, M. J., & Kalinowski, M., "On the interaction between software engineers and data scientists when building machine learning-enabled systems," in Bludau, P., Ramler, R., Winkler, D., Bergsmann, J. (eds) Software Quality as a Foundation for Security. SWQD 2024. Lecture Notes in Business Information Processing, Springer, Cham., vol 505, 2024, https://doi.org/10.1007/978-3-031-56281-5_4.

[57] Ananthi, S., B L., M S. C., and S S., "Framework for platform independent machine learning (ML) model execution," 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), Bengaluru, India, 2024, pp. 728-732, January 2024, doi: https://doi.org/10.1109/IDCIoT59759.2024.10467931.

[58] Mailach, A., and Siegmund, N., "Socio-Technical anti-patterns in building ML-enabled software: insights from leaders on the forefront," 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE), Melbourne, Australia, 2023, pp. 690-702, July 2023, https://doi.org/10.1109/ICSE48619.2023.00067.

[59] Villamizar, H. Kalinowski, M., "Identifying concerns when specifying machine learning-enabled systems: a perspective-based approach," SBQS '24: Proceedings of the XXIII Brazilian Symposium on Software Quality, Salvador Bahia Brazil, pp. 673 – 675, December 2024, https://dl.acm.org/doi/full/10.1145/3701625.3701696.

[60] Joshi, A., "MLOps mastery: streamlining machine learning lifecycle management," International Journal of Science and Research, vol. 13, iss. 1, pp. 1807-1815, January 2024, https://doi.org/10.21275/sr24628132316.

[61] Rella, B. P. R., "MLOPs and DataOps integration for scalable machine learning deployment," International Journal for Multidisciplinary Research, Vol. 4, Iss. 1, pp. 1-20, February 2022, https://www.ijfmr.com/papers/2022/1/39278.pdf.

[62] Soh, J., Singh, P., "Machine learning operations," In: Data Science Solutions on Azure. Apress, Berkeley, CA., pp. 259-279, December 2020, https://doi.org/10.1007/978-1-4842-6405-8_8.

[63] Burgueño-Romero A. M., Benítez-Hidalgo A., Barba-González C. and Aldana-Montes J. F., "Toward an open source MLOps architecture," in IEEE Software, vol. 42, no. 1, pp. 59-64, Jan.-Feb. 2025, https://ieeexplore.ieee.org/document/10588954.

[64] Bayram, F., & Ahmed, B. S., "Towards trustworthy machine learning in production: an overview of the robustness in MLOps Approach," ACM Computing Surveys, vol. 57, iss. 5, article 121, pp. 1-35, January 2024 https://doi.org/10.1145/3708497.

[65] Symeonidis G., Nerantzis E., Kazakis A., and Papakostas G. A., "MLOps - definitions, tools and challenges," 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 2022, pp. 0453-0460, March 2022, doi: https://doi.org/10.1109/CCWC54503.2022.9720902.

[66] Yashwanth Sai Krishna, M., & Gawre, S. K., "MLOps for enhancing the accuracy of machine learning models using DevOps, continuous integration, and continuous deployment," Research Reports on Computer Science, Special Issue, pp. 97–103, June 2023, https://doi.org/10.37256/rrcs.2320232644.

[67] Booth, J., Metz, D.W., Tarkhanyan, D.A., Cheruvu, S., "Machine learning security and trustworthiness," In: Demystifying Intelligent Multimode Security Systems. Apress, Berkeley, CA., pp. 137-222, July 2023, https://doi.org/10.1007/978-1-4842-8297-7_5.

[68] Chen, H., and Babar, M. A., "Security for machine learning-based software systems: a survey of threats, practices, and challenges," ACM Computing Surveys, vol. 56, iss. 6, Article 151, pp. 1-38, February 2024, https://doi.org/10.1145/3638531.

[69] Shirazi, S. H. A., Naghibijouybari, H., and Abu-Ghazaleh, N., "Securing machine learning architectures and systems," GLSVLSI '20: Proceedings of the 2020 Great Lakes Symposium on VLSI, pp. 499–506, September 2020, https://doi.org/10.1145/3386263.3409104.

[70] Chittibala, D. R., and Jabbireddy, S. R., "Security in machine learning (ML) workflows," International Journal of Computing and Engineering, vol. 5, no. 1, 2024, https://doi.org/10.47941/ijce.1714.

[71] Bhuvan, S., "A study on governance framework for AI and ML systems," ShodhKosh: Journal of Visual and Performing Arts, vol. 4, no. 2., December 2023, https://doi.org/10.29121/shodhkosh.v4.i2.2023.1923.

[72] Alsagheer D., Xu L., Shi W., "Decentralized machine learning governance: overview, opportunities, and challenges," in IEEE Access, vol. 11, pp. 96718-96732, September 2023,https://doi.org/10.1109/ACCESS.2023.3311713.

[73] Zhang, X., and Jaskolka, J., "Conceptualizing the secure machine learning operations (SecMLOps) paradigm," 2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS), Guangzhou, China, 2022, pp. 127-138, March 2023, https://doi.org/10.1109/QRS57517.2022.00023.

Bridging Machine-Readable Code of Regulations and its Application on Generative AI: A Survey

Samira Yeasmin, Bader Alshemaimri
Software Engineering Department, King Saud University, Riyadh, Saudi Arabia

***Abstract*—Machine-Readable Code (MRC) and Machine-Readable Regulations (MRR) enable the conversion of complex regulations into structured formats such as JSON, XML, and X2RL, allowing machines to parse and interpret regulatory texts efficiently. Currently, organizations face challenges in regulatory compliance due to the complexity of regulations, frequent updates, and difficulty in identifying changes that impact policies and procedures. Existing literature provides guidance to a certain extent on how to anticipate regulatory modifications or ensure timely compliance. This review examines current literature on applying machine learning (ML) and Generative AI (GenAI) to extract, structure, and interpret regulatory content. It surveys techniques for converting regulations into machine-readable formats, predicting regulatory changes, and assessing alignment with real-world modifications issued by regulatory bodies. The findings indicate that using MRC, MRR, and AI enables automated compliance checks, faster detection of violations or errors, standardized compliance processes, real-time monitoring, and automatic report generation. These approaches can significantly enhance regulatory adherence across industries, particularly in sectors such as finance, where compliance is critical.**

***Keywords*—Regulatory compliance; natural language processing; machine learning; machine-readable code; Machine-Readable Regulations; generative AI; large language models; RegTech; conflicting regulations; regulation issuance**

I. Introduction

While regulations are an essential part of modern society, it is important to follow the process of regulatory change management. Regulations help to protect people and the environment, but they can be complex and difficult to understand. The complexity of regulations makes it challenging for businesses and individuals to comply with; therefore, recent studies have shown a vast amount of contribution to this field of science, as discussed in section 4. One way of ensuring regulatory compliance is to convert the regulations into machine-readable code (MRC) that can be easily understood and processed by computers.

Once converted, they can be used to train Large Language Models (LLMs) to understand and respond to them. LLMs work with massive amounts of text data and have the capability to understand and generate human language. On the basis of a given input, it has the ability to predict the likelihood of word sequences or generate new text. The larger the dataset to train this neural network, the more accurate the outcomes will be. It can be trained on large datasets of text, such as laws, regulations, and other legal documents. Given the capabilities of LLMs, an AI tool can be developed to highlight similarities between new policies and outdated regulations to help people understand the areas of change. The tool could also be used to identify the entity or organization that issued each regulation, which can help understand the context of the regulation and its applicability to a particular situation. This can allow the regulatory industry to check for compliance with regulations automatically.

Developing a tool based on fine-tuned LLMs to highlight similarities between regulations and updated policies to identify areas of change for the issuing entity or organization is a challenging task, but it has the potential to be a valuable tool for businesses and individuals who need to comply with regulations. Given that the process of regulatory change management requires a significant amount of time because of the complexity of the regulations, the long regulations, and the amount of text that requires investigation to identify change requires a significant amount of manual work and effort. A generative AI can generate new and meaningful content, including text, images, and audio, based on a huge amount of training data. With that being said and noting the capabilities of GenAI, a GenAI-based tool can make a significant impact on regulatory change management, contributing to reducing the time and effort needed to update the regulations and make the process more efficient.

Regulation complexity and understanding it are a crucial problem, as evident in the Global Financial Crisis in 2008 [1]. RegTech has since been introduced. However, RegTech alone is insufficient to ensure that regulations are followed and implemented correctly. Moreover, to ensure that regulations are applied in the respective industries with full implementation, it is important to identify all regulations and legislation in the entities that are updated to match the regulatory bodies.

While LLMs are being used in several different industries, they hold an important position in making a great impact on regulatory change management, contributing to reducing the time it takes to update policies and identifying outdated regulations to ensure compliance with current standards and regulations for organizations. As mentioned above, firms must update their policies every year, with a significant amount of time invested in this process. An AI model to complete the same task will help immensely, not only in highlighting outdated sections of the regulations but also in reducing the time and effort it takes to reflect the updates.

It is important to mention that, as the main sources, IEEE, ACM, ScienceDirect, and SpringerLink were chosen. When looking for primary studies, backwards snowballing was used

to find and scan the references of articles that were relevant to the research.

With RegTech and the emerging implications of machine learning, such as generative AI and Machine-Readable Regulations, in the world of legislation and legality, this study focuses on a thorough analysis of Machine-Readable Regulations and their application to machine learning as well as natural language processing, with a focus on generative AI.

This study is structured as follows: Section II describes the objective and contribution of the study in this field of regulatory change management and the significance and potential of ML models, followed by Section III, which focuses on the background study of regulations and compliance and discusses the challenges in interpreting them. It also presents different techniques that have been implemented to convert complex regulations into machine-readable code. Section IV describes in detail the current implications of machine-readable codes of regulations and their benefits and challenges industry-wise. Different techniques for extracting Machine-Readable Regulations and applications of machine-readable codes of regulations are presented in Section V and Section VI. Finally, Section VII presents the conclusion of the survey analysis, and Section VIII discusses the challenges and future directions of machine-readable codes of regulations, offering a guide to researchers.

II. Objectives and Contribution

Regulations are significant business processes that every business and individual must follow to be compliant with regulatory bodies. Regulations and legislative laws can be difficult to understand, although they are written in human language. Not every individual can understand the legal terms and conditions and, therefore, might face difficulty in following them. Again, businesses must be compliant with regulatory bodies to ensure firms are following the changes and updates required to protect rights and the environment and be accountable for their actions. While it takes a significant amount of time to understand the regulations, find duplicates or changes in regulations and update internal policies accordingly, firms and businesses hire many compliance and legal experts to manage these mountains of regulatory data. Therefore, it becomes crucial to automate this time-consuming task. Converting the regulations and legislative laws written in natural language into MRC to identify hidden patterns and duplicates to highlight conflicting regulations can accelerate digitization. ML's influence can play a significant role in reshaping risk and compliance management. An AI model based on fine-tuned LLMs that can highlight and identify hidden patterns and duplicates in conflicting regulations to underline changes and allow businesses and individuals to be up to date with policies can be valuable implementation in the regulatory industry.

This study aims to identify and study existing and emerging ML models to achieve the following objectives:

- Anticipate studying the hidden patterns between the regulatory changes and the outdated policies where changes must be applied.
- Dig deep into the historical patterns to forecast/predict future changes in regulatory compliance with the use of ML models.
- Study the implementation of machine learning to correctly identify the prediction of regulatory changes and correlate them with the actual changes applied by the regulatory bodies.
- Investigate the correlation between the regulatory changes and the impact on the businesses and individuals who must be compliant with the regulations and legislative laws.

III. Background

A. Regulations and Compliance

Regulations are significant business processes that every business and individual must follow to comply with regulatory bodies. Regulations and legislative laws can be difficult to understand, although they are written in human language. Not every individual can understand the legal terms and conditions and, therefore, might face difficulty in following them. Again, businesses must be compliant with regulatory bodies to ensure that firms are following the changes and updates required to protect rights and the environment, and be accountable for the actions that they perform. While it takes a significant amount of time to understand the regulations, find duplicates or changes in regulations, and update internal policies accordingly, firms and businesses hire many compliance and legal experts to manage these mountains of regulatory data. Therefore, it becomes crucial to automate this time-consuming task. Converting regulations and legislative laws written in natural language into MRC to identify hidden patterns and duplicates to highlight conflicting regulations can accelerate digitization. ML influence can play a significant role in reshaping risk and compliance management.

Aligning with national or international regulatory bodies in respective jurisdictions or industries is regulatory compliance. These regulations are internal and involve processes and procedures aimed at streamlining internal business requirements. These regulations not only help in internal processes but also ensure adherence to laws, regulations, guidelines, and specifications relevant to the respective industries. Some examples of regulatory compliance include the Payment Card Industry Data Security Standard (PCI DSS), the Health Insurance Portability and Accountability Act (HIPAA), the Federal Information Security Management Act (FISMA), the Sarbanes-Oxley Act (SOX), the EU's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). As the number of rules increases and the techniques for breaching them increase, many new laws and regulations are developed and introduced so that violations and fraud can be lessened. That said, firms or businesses must also update their old and new regulations and policies when a new law or regulation is passed. This is a manual process that requires the regulatory change management process to be streamlined. Regulatory change management ensures that firms comply with the changes made by the regulatory bodies in the regulations. This activity includes keeping up with the regulatory content,

understanding all the regulations that apply to them, finding out what policies to update and how, etc. This is a complex and time-consuming task for which financial technology (fintech) and financial institutions (FIs) hire many compliance/legal experts to manage mountains of regulatory data.

Regulatory Technology (RegTech) is a technological solution for regulatory compliance and regulations. The main objective of RegTech is to improve regulatory processes. It helps with preparing reports on new and old regulations. One of the objectives of RegTech is to reduce the amount of work completed for manual reporting and to digitize the process. In addition, RegTech helps with risk assessment, identity management and control, transaction monitoring, and fraud prevention, such as money laundering. Furthermore, they address risk assessment [2], [4]–[6]. These are some of the activities that are conducted to ensure regulatory compliance. RegTech aims not to change the industry in which it is implemented but rather to ease the process of regulatory compliance by helping large institutions with the regulatory burden, compliance, and sudden or enormous changes in regulations, and keeping the policies up to date. Regulations usually operate under constant uncertainty, and it becomes more difficult to identify what should be updated to ensure regulatory compliance. This is where RegTech plays a role. FinTech and FIs have started using RegTech to allow themselves to comply with the regulations of regulatory bodies. However, to accelerate digitization, human involvement and interference in RegTech must be minimized, especially in regard to managing regulatory changes.

B. Challenges in Interpreting and Complying with Regulations

As mentioned above, a significant problem lies in regard to regulation complexity and accurately understanding it. As evident in the Global Financial Crisis in 2008 [1], the complexity of regulations played a vital role in the crisis. RegTech has since been introduced. However, RegTech alone is insufficient to ensure that regulations are followed and implemented correctly. Moreover, to ensure that regulations are applied in the respective industries with full implementation, it is important to identify all regulations and legislation in the entities that are updated to match the regulatory bodies.

Along with understanding the regulations, identifying the changes being made to the regulations so that the internal policies can be updated is highly important. This part of managing regulatory changes allows firms to comply with the changes made by regulatory bodies. There are several activities related to this, including understanding the regulatory content, identifying the changes, and determining what internal policies must be updated. For a human, this is a complex and time-consuming task. Firms usually hire many compliance and legal experts to manage mountains of regulatory data, which also becomes an expensive cost.

The number of regulatory changes taking place in organizations is surprisingly large. According to a case study published by iBM [2], the U.S. Code of Federal Regulations (CFRs) contains more than 174,000 pages. It is as long as 25 miles long if printed. The amount of regulatory change has tripled in the last five years, the case study from iBM claims, which was conducted in 2021. This can be a resource-intensive task to understand the complex regulations and take a significant amount of time to reflect on the updates. As a rule, policies should be updated every 13 years [3], [7]. Moreover, high-risk industries such as healthcare, public safety, and financial services are advised to update their regulations more frequently [7], [8]. The time required to update policies depends on multiple factors, including the complexity of the regulations, the scope of the changes needed, the size of the organization, and the number of stakeholders involved. Regular review and updates of regulations are essential to ensure compliance, effectiveness, and alignment with the evolving needs of the business and the external environment.

The potential risks of not updating outdated regulations are significant and can impact various aspects of an organization. It can lead to noncompliance with current standards and regulations [9], [10], security vulnerabilities [11], increased operational costs [9], a lack of clarity and consistency [11], obsolete technology and manual workarounds.

Large Language Models (LLMs) can be used to identify similarities between outdated regulations and highlight the changes needed to match new regulations published by regulatory bodies. This process involves comparing the language and content of old and new regulations and understanding the implications of these changes for various stakeholders. LLMs can play a crucial role in identifying similarities between outdated regulations and highlighting the changes needed to match new regulations published by regulatory bodies. This can help ensure that regulatory compliance processes remain accurate and efficient in the face of rapid technological advancements and changes in regulatory landscapes.

While LLMs are used in several different industries, they hold a prominent position in making a great impact on regulatory change management, contributing to reducing the time it takes to update policies and identifying outdated regulations to ensure compliance with current standards and regulations for organizations. As mentioned above, firms must update their policies every year, with a significant amount of time invested in this process. An AI tool to complete the same task will help immensely, not only in highlighting outdated sections of the regulations but also in reducing the time and effort it takes to reflect the updates.

C. Machine-Readable Data

While machine learning has shown numerous applications in several different industries, RegTech is no different. Businesses and individuals have started using RegTech to allow themselves to comply with the regulations of regulatory bodies. However, to accelerate digitization, human involvement and interference in RegTech must be minimized, especially in regard to managing regulatory changes. It is extremely important to comply with these regulations to speed up the unprecedented customer traffic channeled by businesses. To ensure due diligence, review of customers/businesses (KYC/KYB) and be on top of AML

regulations, businesses and individuals, especially fintech and FIs, are required to conduct periodic checks to establish preventive measures against cyber fraud, money laundering, or terrorist financing activities. This is a complex and time-consuming task for which businesses hire many compliance/legal experts to manage mountains of regulatory data. The influence of ML can play a significant role in reshaping risk and compliance management to accelerate digitization. In addition, regulations can be challenging to understand. Converting them into MRCs and allowing LLMs to be trained and finetuned to achieve a specific specialization can ease the task of complying with regulations. The significant changes in the regulations must be reflected in internal policies for businesses and individuals to follow. It is necessary to ensure consistent governance with regulations.

1) Machine Readable Regulations (MRR): The ability of machines to parse legal documents is what Machine-Readable Regulations are capable of doing. MRR allows access to legal documents and regulations in a machine-readable format. Usually, regulations and policies are drafted and published in digital format, which is easily accessible for humans but not for machines. Machines cannot access legal documents published in digital format. Therefore, it is essential to structure them in a correct form that is accessible to machines, such as XML [1]. In addition, machine-accessible legal documents predefine special delimiters identifying different parts of the document. The term that can be used here is Machine Readable Format (MRF) for documents accessible in this structure, which XML offers. XML can be used to structure any content, not only legal documents. Other examples include the United States Legislative Markup (USLM) used in the US only and the UN's Akoma Ntoso MRF, which can be considered an international MRF [1].

2) Semantic Machine-Readable Regulations (SMRR): MRR focuses on the structure and accessibility of legal documents, which makes the content readable to machines. However, it is important to understand the meaning of the content that Semantic Machine-Readable Regulations (SMRR) allows. The main focus of SMRR is to have semantics for the structured content to make it useful and semantically machine-readable. This semantics produces internal and external insights. Internal insight allows us to understand various possible meanings of the content itself. However, external insight explains the relationship between the provisions of the document and its corpus. These insights provide metadata that machines can access. X2RL and JSON formats are examples of SMRR formats [1].

3) Machine-Executable Regulations (MER) and Machine-Consumable Regulations (MCR): While ensuring that machines have access to regulations is important, it is also important to carry out functions using these metadata. Functions include generating compliance reports, monitoring, and updating regulations. Codifying legal documents is one of the means of achieving the function offered by Machine-Executable Regulations (MER) and Machine-Consumable Regulations (MCR). The codification allows the underlying meaning of the regulations to be converted into a code that machines can execute. Regulations or legislation can be Boolean logic or conditional statements that can be executed with one or several lines of code. MERs and MCRs are extensions of SMRRs. Unlike SMRRs, MERs and MCRs do not produce metadata, but they execute MCR content [1].

4) Rules as Code (RaC)/Legislation as Code (LaC): MRR, SMRR, and MER/MCR are concepts and traits of Rules as Code (RaC) [equally Legislation as Code (LaC)]. The main difference between the previously explained concepts and RaC is integrability. Breaking down the legal corpus into basic legal items and encoding them into basic modules are the first steps of RaC. These modules then accept structured input and produce structured output. The RaC document integrates the set of modules in a constructed document, ensuring that there are no repetitions of legal concepts that can be possible in MER/MCR, since they codify legal concepts only without considering coding repetition, such that some legal items that are common across documents are coded several times separately [1].

D. Generative AI (GenAI)

GenAI is a type of AI that can create new content, such as text, images, and music. It can be used to generate realistic synthetic data, which can be used to train other AI models or to generate new products and services. In the context of regulatory compliance, GenAI could be used to generate realistic synthetic datasets of regulations, which could be used to train AI models to identify conflicting regulations and gaps in regulations. On the basis of several types of inputs, Generative AI (GenAI) creates new content. This content can include text, images, sounds, animations, 3D models, etc. GenAI uses Neural Networks (NN) to identify hidden patterns and structures within any type of data and generate new content. The training of GenAI is not bound to any one of the unsupervised or semi-supervised learning approaches. However, it can be learned and trained in any way. Therefore, organizations can leverage GenAI's features, as it allows them to create foundation models because of the large amounts of unlabeled data. Given that foundation models can be versatile, they are AI-based systems or services whose usage is decided in which industry it is implemented. Some examples of GenAI include GPT-3 and Stable Diffusion [12].

There are five algorithms that provide the framework to develop the GenAI system [13].

1) Overview of generative AI techniques

a) Autoencoders: Autoencoders are a type of neural network that does not require training with prelabeled data [14]. The main use of autoencoders is to compress high-dimensional data into lower-dimensional data. This is called a latent space. The encoder and encoder components are jointly trained. However, because of the low feasibility of AE designs, the change in the latent variables is not high. This is where the Variational Autoencoder (VAE) is introduced, allowing more constraints to be added to latent variables for the desired distribution. While VAEs recognize problems with low feasibility, they produce blurry outputs [13].

b) Generative Adversarial Networks (GAN): A GAN is a deep learning architecture that trains two neural networks to produce authentic data from a given training dataset [15]. The two different networks compete with each other. One network modifies the input data sample as much as possible to produce new data. The other network predicts whether the generated new data belongs to the original input data. The system continues generating new and improved versions of data until it is no longer related to the original input data.

c) Autoregressive model: With the exceptional capability of density estimation, Autoregressive Models sequentially predict each variable component on the basis of prior elements [16]. The performance is affected by the slower processing of a high volume of data. One of the advancements of autoregressive models is the integration of self-attention mechanisms, especially transformer models. This allows parallel processing and selective focus on sequence elements.

d) Diffusion and flow-based model: Differences in operational mechanism, diffusion, and flow-based models are quite similar and often grouped. Diffusion models add noise to data and conduct reverse engineering for data reconstruction [16]. It consists of forward and backward diffusion processes where forward noise is introduced and backward diffusion diminishes this noise. Furthermore, flow-based models leverage normalizing flow to generate a probability distribution between the data and the latent space. This is particularly useful for density estimation.

e) Foundation model (FM): FMs are foundational models that are trained on vast amounts of data and can produce highly accurate and diverse outputs [16]. This can range from images to texts to complex simulations. They can generate very specific tasks on the basis of limited datasets for improved performance. Some examples include image and text generation, multimodal models, and reinforcement learning models.

IV. Machine-Readable Code of Regulations (MCR)

Machine-readable code (MCR) has become a booming field of research in which the focus is mainly centered on how any text can be converted in such a way that it can be easily understood, processed, and parsed by a computer. In regard to MCRs, one area that has gained the attention of regulators is the use of MCRs in regulations. It becomes quite difficult for industries such as FinTech, Financial Institutes, Healthcare, etc., to keep up with the updates and changes in regulations by regulatory bodies and ensure that they are compliant with all the policies and procedures. Machine learning (ML) can help computers process these regulations and make this tedious task easier for the respective compliance departments of different industries.

Machine learning (ML) can add significant value to the regulatory industry in regard to abiding by regulations and ensuring that other businesses and individuals are compliant with existing and new regulations. For example, programs using ML algorithms are already being implemented in the field to detect fraud and anti-money laundering activities or monitor transactions in the financial industry [17]. The ML is trained such that it can read Machine Readable Code (MRC), which has been encoded from the natural language [17]. Furthermore, converting regulations written in natural language is one of the main activities allowing an ML to be trained on this data to support regulatory compliance. Researchers have suggested and offered different approaches to convert natural language into MRC. Some of the prominent forms of machine-readable formats for regulatory documents are XML, JSON and X2RL. The process of converting to a machine-readable format is important to extract the correct meaning of the regulations, which might sometimes follow inconsistent language. In [18], the authors introduced a new machine-readable format and developed X2RL to enhance the semantic representation of regulatory information. The study shows that X2RL helps in enabling a more granular and precise representation of regulatory documents and capturing the underlying semantics. It also discusses how semantic representation can aid in automated interpretation and analysis of regulations, which is a requirement for the use of ML algorithms for regulations. Moreover, [19] introduces a "legislative recipe" as the idea of a structured syntax designed to represent legislative texts in a machine-readable format. This design includes provisions, definitions, and references that enable precise and standardized representation of the legislative components. It also accommodates the semantic annotation and hierarchical nature of the regulations. In a previous paper [20], the authors described the foundational concept of the JSON schema and how it can be a tool for describing and validating the structure of JSON data. It helps specify the structure, constraints, and validation rules for JSON documents, which leads to the validation of common types of strings, numbers, and booleans. In [12], the authors aims to improve the encoding of copyright legislation into machine-executable code. To ensure accuracy in "Rules as Code", it is important to involve legally trained individuals. This study revealed that collaborative agreement on key legal terms leads to better-encoded rules. However, different interpretive choices can still occur due to complexities in statutory interpretation and coding language functionality. The authors suggest separating technical validation and legal alignment processes to improve accuracy. The authors used Defeasible Deontic Logic (DDL) and Turnip to convert selected provisions of the Copyright Act into machine-executable code. Furthermore, [21] discussed the inherent complexity of regulatory texts, which are often filled with legal jargon, ambiguities, and references to other regulations. They emphasize the challenges this complexity poses for both human understanding and machine automation. The authors delve into how natural language processing (NLP) and machine learning techniques can be employed to interpret regulatory texts. The authors further emphasized the point that rulemaking allows the automation of compliance, reporting, and enforcement processes with the help of ML technologies.

The research field is booming with several different types of approaches to achieve the goal of an efficient and faster process of converting regulations into MRCs. While it is challenging to convert natural language into machine-readable code, several machine-learning algorithms or techniques have been developed to aid in MRC. The task of converting into MRC falls under the domain of Natural Language Processing

(NLP) and Natural Language Understanding (NLU). Several machine learning algorithms and techniques can be employed for this purpose, including Recurrent Neural Networks (RNNs) [22], Transformer Models [23], Seq2Seq Models [24], Graph Neural Networks (GNNs) [25], and Large Language Models (LLMs) [26]. The authors of [22] reported that an RNN can be used to recognize written text and convert it into MRC. They used the fundamentals of Long Short Term Memory (LSTM) to capture the long-range sequence of handwritten elements to identify noise in handwritten data and segments, and isolate it into individual characters. This study also presents the implementation of Seq2Seq models where the input is the sequence of handwritten texts and produces corresponding text sequences, but with MRC as the output. Moreover, [23] highlighted the importance of an ML technique in understanding software code and generating related and useful comments for specific code snippets. If an ML technique can convert an MRC into natural language, it is an opportunity for researchers to leverage the benefits of these MLs and produce state-of-the-art results in natural language understanding for the MLs. The main goal and result of the study [23] was to use semantic parsing in the context of programming languages and extract meaningful information from code. The authors presented the "Setransformer" architecture, which was specifically designed for the semantic parsing of code. Leveraging transformer models to capture both the structural and semantic information within the code helps in generating code comments in human-like language. The authors of the paper [24] presented a thorough analysis of how the GNN can be leveraged for natural language processing. Exploring different ways of representing natural language, such as graphs, dependency trees, co-occurrence graphs, and knowledge graphs, this study highlights the advantages of using graph-based representations for capturing semantic relationships in language. The survey covered a wide range of NLP tasks where the GNN has been applied, such as sentiment analysis, named entity recognition (NER), relation extraction, question answering, knowledge graph completion, language modelling, and semantic role labelling.

In practice, recent studies have discussed the significance of machine learning (ML) in regulatory change management, particularly in the context of converting regulations into Machine Readable Code (MRC). The implementation of ML algorithms in regulatory industries, such as fraud detection and anti-money laundering, is highlighted. ML models are trained to read MRCs encoded from natural language regulations. There are various approaches for machine-readable formats, including XML, JSON, and X2RL. The complexity of regulatory texts, filled with legal jargon and ambiguities, is acknowledged, and the use of natural language processing (NLP) and machine learning techniques for interpretation is discussed.

A. Objectives and Benefits of Adopting MRCs

1) Efficiency Improvement: The automated extraction of key information from regulatory documents saves time and effort for the compliance department, which is engaged in human interpretation, to identify the meaning of the information, diving deep into the regulations and applying them to the workplace. This makes the compliance process more efficient. As mentioned in [27], a significant amount of time is spent on understanding and identifying areas of change in regulatory policies. Given that MRC can read and process regulations without the involvement of a human resource, it not only saves time but also makes it easier for humans to interpret regulations and understand complex clauses, which must be updated every one to three years [3], [7].

2) Accuracy enhancement: Legally trained individuals are required to ensure the accuracy of policies and regulatory documents. Additionally, each individual interprets the policies in different interpretive choices, which creates statutory interpretation and coding language functionality. The authors [21] suggested that NLP can increase accuracy and eliminate the complexity of understanding regulations by converting them to MRCs. The authors further emphasized the point that rulemaking allows the automation of compliance, reporting, and enforcement processes with the help of ML technologies. The MRC helps minimize the risk of human error, ensuring that compliance activities are executed accurately and in strict accordance with regulatory requirements.

3) Real-time monitoring and adherence: MRC enables monitoring of regulatory changes and allows identification of the areas where the changes have been reflected. The objective of this ability of MRC is to create consistency between the policies implemented in the respective fields and ensure adherence to the regulatory updates by the regulatory bodies. This allows us to adapt to modifications to regulations and maintain adherence.

4) Standardization of compliance processes: A uniform approach is required to be compliant with various jurisdictions and sectors in which the respective compliance department individuals invest their resources, effort, and time. MRC promotes the use of standardized formats and structures for regulations, making it easier to create and understand complex regulations.

5) Automation of reporting: With MRC, an automated process of reporting can be introduced by offering a standardized compliance process, identifying accurate data, and interpreting regulations via MRC. Reporting requires a significant amount of data analysis to abide by regulations accurately. The MRC can ensure adherence to regulations and policies reflecting compliance reporting.

B. Evolution of Machine-Readable Regulations (MRRs) and Current Development [1]

Access to data, meaning that converting data in a standardized form that computers can process, is one of the cornerstones and requirements of MCRs. Computers with the ability to parse legislative content and regulations are quite challenging. Usually, regulations are written in electronic formats, such as MS Word and PDF, which are easy for the human eye to read but not easy for machines to process and identify the meaning of the regulations. To address the challenges of machines being able to access legal content, the

MRR was formed. The readability of regulations for machines is affected by the structure of the regulations or the legal contents. This means that the format of the content must be in a way that helps to discern the components so that the identification of the components is similar to how a human does. This structure is referred to as Machine Readable Format (MRF), and one example of such a format is XML. It can be used to structure any content, not only legalese or regulations. MRF is the first of its kind that helps in transforming legal content into a form that can be processed and parsed by machines. This is an essential step toward RaC. While MRR focuses on the structure of regulatory documents, SMRR focuses on the enrichment of texts as well. The process of SMRR includes the injection of semantics in the machine-readable structure in the form of metadata. This is similar to how a human interprets a document and drafts the semantics. To further improve the process, codification has been added to the process of parsing regulations by machines. MER and MCR allow converting the documents into Boolean logic and further converting it to computer code, which is called codification. To conduct the process of codification, the text must be structured and enriched, which is also part of the MER and MCR. To ensure the integrability of the documents, the document is broken into a corpus that consists of basic rules and modules. These different modules are then converted into machine-readable code, ensuring that the same meaning of different modules is not repeated during the parsing process. This process is called RaC and LaC.

C. Common Standards and Machine-Readable Formats

1) JavaScript Object Notation (JSON): The JSON schema provides a standardized schema language for JSON documents [20]. It allows developers to specify the structure of JSON documents, constrain their content, and verify the integrity of requests and responses in APIs. The JSON schema is used for various purposes, including data validation, document-oriented databases, automated document generation, and schema validation tools. JSON is commonly used as a machine-readable format since it is lightweight and easy to parse for machines. It is ideal for transmitting data between systems.

2) XML: XML is one of the most prominent machine-readable formats [1], [18]. The US's regulations, named the United States Legislative Markup (USLM), and the internationally recognized regulation, called the UN's Akoma Ntoso, are XML-based. These XML-based regulations contain a hierarchy of tags and attributes. It helps improve the internal structure of the legal document rather than the content. They sufficiently lack the developed methods of connecting the bodies of legal documents in a machine-readable way. This helps reduce the management cost of legal documents; however, the main goal is to minimize the economic cost of legislation for those who must comply with the body of the regulations.

3) eXtensible Regulatory Reporting Language (X2RL): Authors in [18] proposed a new machine-readable format that, along with having tags and attributes such as XML, has additional new attributes, tags, and models designed to enrich the metadata of the legal document's content. According to the authors, X2RL helps in processing the legal content, intent, scope, and meaning of the legislation and regulations. In addition, it identifies the depth of the external structure of the documents to link multiple documents together, which reduces the economic cost and allows efficient data interchange. The authors introduced <**provision**> as the basic container for any document fragment. Strengthening the external structure of the legislation and expanding the ontology to include new elements to be able to describe the legal contents in detail are two more goals X2RL achieves.

4) YAML Ain't Markup Language (YAML): YAML is also a markup language used to provide a structure for data [28]. Like JSON, YAML stores data in a human-readable format by providing serialization and offering text marking. YAML can be used in place of JSON since it is now a subset of JSON. Supporting scalar datatypes such as lists and arrays, it stores data in a text file.

5) Resource Description Framework (RDF): RDF is mainly used to convert data on the web in a machine-readable format [29]. In RDF, information is represented in the form of triples, which consist of subject-predicate-object statements. The subject and object represent resources, whereas the predicate indicates the relationship between them. Facilitating data integration and interoperability, RDF is commonly used to create linked data.

6) Web Ontology Language (OWL): Like RDF, OWL is used for data on the web. It represents domain knowledge via classes, properties, axioms, and instances for use in a distributed environment [30]. Representing a semantic representation of domain knowledge, OWL supports efficient reasoning and expressive power. The authors of [31] proposed a new method of generating OWL ontologies from XML data sources to highlight how XML data can be extracted to represent domain knowledge in OWL format.

D. MRR Use Cases

Converting legal content and regulations into machine-readable language has made strides in the RegTech industry. It is a transformative movement for the global regulatory community to ease the complexity of regulations, legislation, and policies, allowing entities to assist in complying with new regulations and drafting or updating old legislation and policies. RegTech can be further improved with the essence of MRR so that machines can analyze legal content to parse the content itself. The following are some of the use cases that are framed by MRRs and offer improvements in the regulatory industry.

1) Committed to regulations: MRR adaptation is effective and can be an efficient way of ensuring that an entity is committed to following regulations. Regulators can easily identify if entities are misinterpreting regulations and making erroneous decisions. MRR allows entities to refrain from making any judgment errors. It improves the quality of compliance. An MRR implementation helps in the use of machine learning algorithms to study the regulations and

determine whether they are correctly implemented and followed within the organization.

2) Regulations updates: MRR does not merely translate the regulations into machine-readable language but also creates the opportunity to use machine learning algorithms to identify the updates in regulatory bodies and how existing regulatory policies must be updated. It helps highlight outdated rules and update them with new implementations. Financial institutes can prepare MRR-based handbooks and search places where humans and machines can search for specific rules and regulations. MRR can also help in identifying duplicate policies within an organization and enhancing them by removing duplicates and replacing them with the correct format. This signifies and symbolizes the movement of machine-analyzable legal content and the process of generating legal documents.

3) Implementation of principle-based regulations: To facilitate measurements of the competency of licensing requirements for financial institutes and ensure that they meet supervisory expectations, following regulations is important. MRR allows the implementation of regulations to meet the requirements of the regulatory body by aggregating relevant documents for analysis and comparing them with what is being followed within the organization/entity. A principle-based regulation requires a greater level of understanding by machines, which is needed for the compliance of financial institutions with regulatory bodies. An extensive MRR adaptation in the financial sector to meet licensing requirements is a significant and effective approach.

4) Financial reporting: With the help of MRR, the reporting process can be much more efficient and faster. Using technology, the current reporting process can be automated to be more accurate and consistent by standardizing the description and identification of data, as well as digitizing reporting instructions and generation through MRR. Reporting is crucial in regard to abiding by regulations in the right way. The MRR can ensure 100% adherence to regulations and policies and reflect the same in financial statements and compliance reporting.

V. Techniques for Extracting Machine-Readable Regulations

MRR allows machines to parse complex legislation, regulations, and policies that humans might find difficult to interpret. It not only converts regulations into machine-readable formats but also offers the chance to identify areas of change, ensuring the compliance of policies and rules that a body is following with regulatory bodies. NLP offers extensively utilized techniques to extract and parse regulations from textual documents in various domains, such as legal, compliance, governance, and policymaking. As mentioned earlier, this practice of converting regulations into machine-readable formats involves identifying and understanding rules, interpreting them for machines matching the human meaning, and allowing different sectors to follow the regulations properly. There are multiple techniques to extract text from regulations so that legislation can be considered Machine-Readable Regulations.

1) Text processing: Texts must undergo preprocessing steps such as tokenization [31], stop word removal [32], and stemming or lemmatization to clean and normalize the text. To conduct NLP tasks accurately and efficiently, it is crucial to analyze and process texts.

2) Named Entity Recognition (NER): Once the text is preprocessed, there are multiple ways of extracting regulations via NLP. One of them is Named Entity Recognition (NER) [33]. It involves identifying and categorizing entities mentioned in the text into predefined categories such as organization names, dates, locations, and, most importantly, regulatory entities such as laws, rules, and directives. According to [33], NER focuses on recognizing particular designators from texts. It requires predefined semantics such as person, location, and organization. It not only extracts information but also plays an essential role in applications such as text understanding, information retrieval, automatic text summarization, question answering, machine translation, and knowledge base construction. There are several techniques applied in NER, including rule-based approaches, unsupervised learning approaches, feature-based supervised learning approaches, and deep learning-based approaches.

3) Rule-based approaches: Rule-based approaches do not need any annotated data. They are based on handcrafted rules [34]. The rules are designed on the basis of domain-specific semantics and syntactic–lexical patterns. They capture specific patterns or formats of regulations within the text. Since this approach focuses on a specific domain, knowledge becomes difficult to transfer to a different domain system.

4) Topic modelling: Topic modelling is an analytical tool used to evaluate data [34]. It focuses on the main topics or themes of the textual contents and helps identify them. It becomes easier to understand and extract themes of the topics of the regulatory documents, and make it easier to determine the meaning of each. Some of the topic modelling approaches include Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF).

5) Text classification: As the name suggests, text classification is a model trained to categorize textual documents into different domains or types. This approach might seem similar to topic modelling, but text classification focuses on the broader area of the regulations, segregating them into regulations for each sector where the regulation is applied [35]. Examples include financial regulations, environmental regulations, and healthcare regulations. Supervised learning algorithms, such as Support Vector Machines (SVMs) or deep learning architectures like Convolutional Neural Networks (CNNs), can be utilized for text classification.

6) Semantic analysis: Another NLP approach for Machine-Readable Regulations is semantic analysis. It helps in understanding the meaning and context of regulatory text [36]. It focuses on the relationships between words and

phrases to interpret and infer the intent and implications of regulations.

7) Information extraction: While it becomes difficult to extract the meaning of complex regulations, information extraction approaches are applied to extract structured information from unstructured regulatory text [37]. The extraction process involves identifying key elements such as obligations, permissions, prohibitions, and conditions mentioned in the regulations. The entire process includes sentence segmentation, tokenization, part-of-speech tagging, entity reorganization, entity disambiguation, relation extraction, and event extraction.

8) Knowledge graph: While the information extraction approach focuses on extracting a structured format of regulations from unstructured regulations, knowledge graphs are constructed to represent the relationship between the structured and interconnected formats of extracted texts [38]. The extracted entities are organized in a graph-like structure that represents the relationships and attributes of the regulations.

TABLE I. TECHNIQUES FOR EXTRACTING MACHINE-READABLE REGULATIONS (MRR)

Technique	Purpose	Methodology	Advantages	Review
Text Processing [31], [32]	Cleans and normalizes text for NLP tasks.	Includes tokenization, stop-word removal, stemming, and lemmatization.	Improves accuracy and efficiency of downstream NLP techniques.	Requires domain adaptation to handle legal language nuances [39].
Named Entity Recognition (NER) [33]	Identifies and classifies key entities in regulatory text (laws, rules, organizations, etc.).	Uses predefined semantic categories (e.g., person, organization, law). Methods include rule-based, unsupervised, supervised, and deep learning approaches.	Enables extraction of relevant regulatory entities for structured representation.	Rely on predefined semantics.
Rule-Based Approaches [34]	Extracts regulatory patterns using handcrafted rules and domain-specific semantics.	Relies on syntactic and lexical patterns defined by experts.	Effective for well-defined, narrow domains with consistent language.	Approach focuses on a specific domain; knowledge becomes difficult to transfer to a different domain system.
Topic Modelling [34]	Identifies hidden themes and major topics in regulatory documents.	Uses algorithms such as LDA and NMF.	Helps understand and categorize large volumes of regulatory text by themes.	May not capture context or subtle legal relationships; requires post-interpretation [40].
Text Classification [35]	Categorizes regulations into domains (e.g., financial, environmental, healthcare).	Employ supervised learning models such as SVM or deep learning (CNNs).	Enables domain-specific regulation management and retrieval.	Requires labeled datasets as they categorized by sector
Semantic Analysis [36]	Interprets the meaning and intent of regulations.	Analyzes relationships between words and phrases to infer implications.	Provides deeper understanding of regulatory intent; improves contextual interpretation.	Complex to implement; dependent on advanced linguistic and contextual models [41].
Information Extraction (IE) [37]	Convert unstructured regulatory text into structured data.	Involves segmentation, tagging, entity recognition, disambiguation, relation, and event extraction.	Identifies obligations, permissions, and conditions for compliance automation.	Challenging for long, nested legal clauses requires hybrid rule-based and ML methods [42].
Knowledge Graphs [38]	Represents structured relationships among extracted regulatory entities.	Organizes entities and relationships into interconnected graph structures.	Enables reasoning, querying, and visualization of regulatory dependencies.	Building and maintaining accurate graphs requires ongoing data validation and ontology design [43].

Table I presents a summary of the different techniques for extracting MRRs. Mainly, these techniques form a layered approach to converting unstructured regulatory texts into structured, machine-readable formats. The several ways of extracting MRRs are to prepare and identify key elements of the text, extract patterns, categorize and organize the data, and lastly, understand meaning, relationships, and build structured representations. They can be grouped as foundational, such as text processing and NER, ensuring clean, labelled data for interpretation. Intermediate, including rule-based extraction, topic modelling, and text classification, providing domain organization and thematic clarity. Finally, more advanced approaches are semantic analysis, information extraction, and knowledge graph, extending these efforts into contextual understanding and structured representation. They establish a pathway for transforming unstructured regulatory texts into interpretable, machine-actionable formats that support automation, consistency, and scalability in compliance management.

VI. Applications of Machine-Readable Code of Regulations in Machine Learning

A. Regulatory Compliance Automation

Through NLP, the extracted texts from regulations providing a structured format of data can be analyzed and processed by machine learning algorithms. ML can identify patterns between historical regulatory data and the current application of regulations to ensure that industries are compliant with regulatory bodies. In addition, ML models can extract relevant information from regulatory documents to generate compliance reports automatically, reducing manual human work and effort. ML can dynamically monitor regulatory changes by keeping itself up to date with regulatory bodies and identifying areas of change in the current implementation of regulations in the respective industry.

B. Risk Assessment and Predictive Analytics

As ML has shown a proven track record of generating predictive outcomes, the application of ML in RegTech is no different. With the help of enough data and training, ML can predict the future development of regulations and assess their potential impact on businesses, providing enough time for regulatory reform. It also allows firms to anticipate compliance risks.

C. Legal Research and Policy Analysis

ML algorithms help summarize complex regulations and extract key insights by analyzing regulatory data with the help of NLP. It offers identifying trends and patterns in the legal industry and aids legal research and policy analysis.

D. Industry-Specific Applications

While RegTech is highly used in the financial industry, its limits do not stop there. In addition to other industries, healthcare and manufacturing are two prominent areas where RegTech is being implemented and has shown great results. In addition to ML algorithms offering automated compliance monitoring, risk assessment, and reporting in the financial sector, ML techniques can help ensure that healthcare industries are compliant with their industry-specific requirements, such as HIPAA in the US. Furthermore, ensuring adherence to safety, environmental, and quality standards, which can be ensured by ML given its impact on RegTech in all aspects of ensuring regulatory compliance, is highly important for the manufacturing industry.

VII. Conclusion

The growing complexity of regulatory frameworks has made compliance both essential and intensive. This study highlights how translating regulatory texts into Machine-Readable Regulations (MRR) can transform compliance from a largely manual, interpretive task into a structured, data-driven process. By enabling machines to interpret and act on regulatory information, MRR offers clear advantages like greater efficiency, consistency, and adaptability in managing compliance across sectors. Beyond automation, integrating Machine-Readable Code (MRC) with AI and knowledge-based systems opens new possibilities for regulation management. Tools like GenAI, ontologies, and federated learning illustrate how compliance can evolve toward predictive, privacy-conscious, and continuously improving systems. Finally, the key insight is that regulatory compliance can no longer rely solely on human interpretation or static systems. The future lies in collaboration between technology and regulatory institutions, where agile frameworks and machine-readable models allow compliance to evolve in step with technological and policy change.

VIII. Challenges and Future Directions

A. Data Quality and Standardization

To ensure that the ML is trained well and that the MRR can generate the expected outcomes, it is important to train the algorithm on complete and consistent data. In addition, a standardized format of training to ensure regulations across different jurisdictions and domains allows the MRR to enhance the quality of its results, promoting the interoperability of data. Advances in semantic technologies, such as ontologies and knowledge graphs, can facilitate the semantic interoperability of Machine-Readable Regulations, enabling more precise and context-aware interpretation.

B. Privacy and Ethical Considerations

The MRR may contain sensitive and confidential information, hindering the privacy of the data. In addition, MRR may exhibit biases leading to unfair outcomes in regulatory compliance decisions on the basis of the training dataset provided to the ML algorithm. Given that the implications of MRR are not limited to a specific industry, it may reveal information affecting the governance framework. Therefore, techniques must be implemented to preserve the privacy of data by utilizing differential privacy and federated learning in MRR. Furthermore, ML algorithms can be developed to indicate bias detection to ensure that fair and ethical guidelines can be introduced in industries and regulatory bodies to preserve the governance framework.

C. Emerging Technologies and Trends

Interdisciplinary collaboration is required so that MRRs can keep up with the emerging changes in the world of technology and ensure that novel regulatory approaches are captured by them. Moreover, regulatory bodies can adopt agile and adaptive regulatory frameworks leveraging MRRs to enable rapid responses to technological advancements and regulatory changes.

Acknowledgment

We acknowledge that a preprint version of this manuscript is available on a non-commercial and academic site.

References

[1] Communications, Space & Technology Commission (CST), "Machine Readable Regulations and Beyond: Tackling Regulatory Complexity," 2023. [Online]. Available: https://www.cst.gov.sa/ar/mediacenter/researchsandstudies/Documents/Tackling_Regulatory_Complexity_Machine_Readable_Regulations_and_Beyond.pdf

[2] Deloitte, "Inside CCO, CISO, CRO, CIA, BOD: Insights from Deloitte," 2017. [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/lu/Documents/about-deloitte/Inside/lu_inside-global-edition-2017-full.pdf

[3] ComplianceBridge Team, "Have You Been Regularly Updating Policies and Procedures?" ComplianceBridge, November 8, 2022. [Online]. Available: https://compliancebridge.com/updating-policies-and-procedures/

[4] D. W. Arner, J. Barberis, and R. P. Buckey, "FinTech, RegTech, and the reconceptualization of financial regulation," Northwestern Journal of International Law & Business, vol. 37, p. 371, 2016.

[5] D. W. Arner, J. Barberis, and R. P. Buckley, "FinTech and RegTech in a Nutshell, and the Future in a Sandbox," CFA Institute Research Foundation, 2017.

[6] D. Yang and M. Li, "Evolutionary approaches and the construction of technology-driven regulations," Emerging Markets Finance and Trade, vol. 54, no. 14, pp. 3256-3271, 2018.

[7] Elliott Davis, "Is it time to review your Policies and Procedures Manual?" Elliott Davis, April 18, 2022. [Online]. Available: https://www.elliottdavis.com/is-it-time-to-review-your-policies-and-procedures-manual/

[8] PowerDMS, "Why it is important to review policies and procedures," PowerDMS Policy Learning Center, December 22, 2020. [Online]. Available: https://www.powerdms.com/policy-learning-center/why-it-is-important-to-review-policies-and-procedures

[9] A. Nadkarni, "How Often Should You Review Your Policies and Procedures?" 24by7Security, 2019. [Online]. Available: https://blog.24by7security.com/how-often-should-you-review-your-policies-and-procedures

[10] C. Anderson, "What are the Risks of Not Updating Policies and Procedures?" Bizmanualz, May 22, 2023. [Online]. Available: https://www.bizmanualz.com/writing-policies-and-procedures/what-are-the-risks-of-not-updating-policies-and-procedures.html

[11] M. Prijic, "Risks of Using Outdated Operating System," IT Convergence, April 6, 2023. [Online]. Available: https://www.itconvergence.com/blog/risks-of-using-outdated-operating-system/

[12] A. Witt, A. Huggins, G. Governatori, and J. Buckley, "Converting copyright legislation into machine-executable code: Interpretation, coding validation and legal alignment," in Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (ICAIL), 2021, pp. 139-148.

[13] T. Sakirin and S. Kusuma, "A Survey of Generative Artificial Intelligence Techniques," Babylonian Journal of Artificial Intelligence, 2023, pp. 10-14.

[14] J. T. Chien, Source Separation and Machine Learning. Academic Press, 2018.

[15] M. Salvaris, D. Dean, and W. H. Tok, Deep Learning with Azure: Building and Deploying Artificial Intelligence Solutions on the Microsoft AI Platform. Apress, 2018.

[16] B. Decardi-Nelson, A. S. Alshehri, A. Ajagekar, and F. You, "Generative AI and Process Systems Engineering: The Next Frontier," arXiv preprint arXiv:2402.10977, 2024.

[17] E. Micheler and A. Whaley, "Regulatory technology: replacing law with computer code," European Business Organization Law Review, vol. 21, pp. 349-377, 2020.

[18] P. A. McLaughlin and W. Stover, "Drafting X2RL: A Semantic Regulatory Machine-Readable Format," MIT Computational Law Report, 2021.

[19] M. Ma and B. Wilson, "The legislative recipe: Syntax for machine-readable legislation," Northwestern Journal of Technology & Intellectual Property, vol. 19, p. 107, 2021.

[20] F. Pezoa, J. L. Reutter, F. Suarez, M. Ugarte, and D. Vrgoč, "Foundations of JSON schema," in Proceedings of the 25th International Conference on World Wide Web (WWW), 2016, pp. 263-273.

[21] J. Mohun and A. Roberts, "Cracking the code: Rulemaking for humans and machines," 2020.

[22] I. Paul, S. Sasirekha, D. R. Vishnu, and K. Surya, "Recognition of handwritten text using long short-term memory (LSTM) recurrent neural network (RNN)," in AIP Conference Proceedings, vol. 2095, no. 1. AIP Publishing, April 2019.

[23] Z. Li et al., "Setransformer: A transformer-based code semantic parser for code comment generation," IEEE Transactions on Reliability, vol. 72, no. 1, pp. 258-273, 2022.

[24] L. Wu et al., "Graph neural networks for natural language processing: A survey," Foundations and Trends in Machine Learning, vol. 16, no. 2, pp. 119-328, 2023.

[25] L. Wu et al., "Graph Neural Networks for Natural Language Processing: A Survey," vol. 16, no. 2, pp. 119–328, Jan. 2023, doi: https://doi.org/10.1561/2200000096.

[26] Nikitas Karanikolas, E. Manga, Nikoletta Samaridi, E. Tousidou, and M. Vassilakopoulos, "Large Language Models versus Natural Language Understanding and Generation," Nov. 2023, doi: https://doi.org/10.1145/3635059.3635104.

[27] "Regulatory Change Management Efficiency, Effectiveness & Agility in Regulatory Change Processes STRATEGY PERSPECTIVE Governance, Risk Management & Compliance Insight," 2021. [Online]. Available: https://www.ibm.com/downloads/cas/MAN6JNY4

[28] A. Loibl, T. Manoharan, and A. Nagarajah, "Procedure for the transfer of standards into machine-actionability," Journal of Advanced Mechanical Design, Systems, and Manufacturing, vol. 14, no. 2, p. JAMDSM0022, 2020.

[29] N. Haider and F. Hossain, "CSV2RDF: Generating RDF data from CSV file using semantic web technologies," Journal of Theoretical and Applied Information Technology, vol. 96, no. 20, pp. 6889-6902, 2018.

[30] G. Antoniou and F. V. Harmelen, F. V. "Web ontology language: Owl," Handbook on ontologies, pp. 91-110, 2009. Berlin, Heidelberg: Springer Berlin Heidelberg.

[31] S. J. Mielke et al., "Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP," arXiv preprint arXiv:2112.10508, 2021.

[32] D. J. Ladani and N. P. Desai, "Stopword Identification and Removal Techniques on TC and IR applications: A Survey," in Proceedings of the 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 466-472, doi: 10.1109/ICACCS48705.2020.9074166.

[33] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 1, pp. 50-70, 2020.

[34] I. Vayansky and S. A. Kumar, "A review of topic modeling methods," Information Systems, vol. 94, p. 101582, 2020.

[35] V. Dogra et al., "A complete process of text classification system using state-of-the-art NLP models," Computational Intelligence and Neuroscience, vol. 2022, 2022.

[36] S. A. Salloum, R. Khan, and K. Shaalan, "A survey of semantic analysis approaches," in Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020), Springer, 2020, pp. 61-70.

[37] S. Singh, "Natural language processing for information extraction," arXiv preprint arXiv:1807.02383, 2018.

[38] I. Mondal, Y. Hou, and C. Jochim, "End-to-end NLP knowledge graph construction," arXiv preprint arXiv:2106.01167, 2021.

[39] D. P. Karanikolas et al., "Strengths and Weaknesses of LLM-Based and Rule-Based NLU Systems," Electronics, vol. 14, no. 15, 2025.

[40] K. Ashihara, "Improving topic modeling through homophily for legal documents," Applied Network Science, vol. 5, 2020.

[41] R. M. Bakker, A. J. Schoevers, van Drie, M. P. Schraagen, and T. de, "Semantic role extraction in law texts: a comparative analysis of language models for legal information extraction," Artificial Intelligence and Law, Mar. 2025, doi: https://doi.org/10.1007/s10506-025-09437-x.

[42] D. Premasiri, T. Ranasinghe, R. Mitkov, et al., "Survey on legal information extraction: current status and open challenges," Knowledge & Information Systems, vol. 67, no. 12, pp. 1–35, 2025.

[43] C. Peng et al., "Knowledge Graphs: Opportunities and Challenges," Artificial Intelligence Review, 2023.

Energy Efficient Workflow Allocation in Cloud Computing Using Improved Grey Wolf Optimization

Md. Mazhar Nezami, Anoop Kumar
Department of Computer Science-College of Computing and Mathematics, Banasthali Vidyapith, Rajasthan, India

***Abstract*—Cloud computing has emerged as a dominant platform for hosting complex applications, offering scalable and flexible resources on demand. However, the dynamic and heterogeneous nature of cloud environments poses significant challenges for efficient workflow scheduling, particularly when aiming to minimize total execution time, energy consumption, and operational cost. In this research, we propose a novel hybrid approach that integrates the Heterogeneous Earliest Finish Time (HEFT) algorithm with an Improved Grey Wolf Optimizer (IGWO) enhanced by differential evolution strategies and survival-of-the-fittest mechanisms. These enhancements strengthen exploration and exploitation by adaptively mutating and refining task allocations while eliminating weaker solutions. The use of HEFT-based initialization provides a strong starting population, and the DE-driven IGWO refinement accelerates convergence and avoids premature stagnation. Together, these two-level optimization strategy ensures faster convergence and higher energy-efficient workflow scheduling compared to earlier HEFT metaheuristic approaches. To evaluate the effectiveness of the proposed hybrid method, extensive experiments were conducted on randomly generated workflows with varying task and dependency complexities. The performance analysis demonstrates that the hybrid HEFT-IGWO approach consistently outperforms standard HEFT, traditional GWO, and standalone metaheuristic techniques in terms of minimizing makespan, reducing energy consumption, and lowering cloud infrastructure costs. This study highlights the potential of combining heuristic initialization with evolutionary optimization to achieve energy-efficient, cost-effective workflow scheduling in cloud computing environments.**

***Keywords*—Cloud computing; energy efficient; workflow; Heterogeneous Earliest Finish Time (HEFT); Grey Wolf Optimization (GWO); makespan; cost**

I. Introduction

Cloud computing has revolutionized the way computational resources are provisioned and consumed, offering scalable, on-demand access to a vast pool of virtualized hardware and software services. It enables organizations to deploy complex applications and workflows without investing heavily in physical infrastructure [1], [2].

A workflow in cloud computing represents a set of interdependent tasks organized to achieve a specific computational objective, such as scientific workflows, big data processing workflows. The workflow scheduling refers to organizing workflow tasks on VMs within the cloud platform, addressing the challenge of balancing energy efficiency with QoS optimization [3].

As cloud environments grow larger and more heterogeneous, efficient workflow allocation, assigning tasks to appropriate virtual machines (VMs), becomes critical. Poor task allocation results in high direct and indirect energy consumption, leading to high costs, reduced reliability, and environmental issues like CO2 emissions [4][5]. If a data center hosts 10,000 servers and each consumes an average of 300 watts, the total energy consumption can reach 3 megawatts per hour. At an average electricity cost of $0.12 per kilowatt-hour, operating such a centre would cost approximately $360 per hour, not including cooling and infrastructure overheads.

Despite significant progress in workflow scheduling, existing algorithms still struggle to maintain a balance between performance efficiency and energy optimization. Traditional heuristics, such as the HEFT algorithm, offer fast scheduling decisions but are limited by their static and local search behaviour. Conversely, metaheuristic algorithms like Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and Grey Wolf Optimizer (GWO) improve global search ability but may converge prematurely or require many iterations.

Therefore, there is a critical need for a hybrid scheduling framework that can combine heuristic efficiency with adaptive global optimization to effectively minimize makespan, energy consumption, and operational cost in dynamic, heterogeneous cloud environments [6].

This study seeks to determine how heuristic and metaheuristic methods can be effectively integrated to enhance workflow scheduling efficiency in cloud computing. It also incorporates differential evolution and survival-of-the-fittest mechanisms within an Improved Grey Wolf Optimizer (IGWO) can strengthen convergence stability and solution diversity. Furthermore, the research investigates the extent to which the proposed hybrid HEFT–IGWO model can reduce makespan, energy consumption, and operational cost compared to traditional scheduling techniques. This research contributes to the advancement of green and sustainable cloud computing by proposing an intelligent hybrid model that improves both computational performance and energy efficiency. The proposed method offers a practical solution for multi-objective workflow allocation, reducing makespan, energy consumption, and cost simultaneously.

The study is designed in the rest of the sections as follows: Section II presents a literature survey. Section III presents the problem definition, problem formulation, and suggested HEFT-IGWO. Section IV describes the detailed algorithmic design, including initialization, mutation, crossover, and survival mechanisms of the HEFT–IGWO approach. Section V provides

the experimental setup, simulation results, and performance analysis, comparing the proposed method with existing approaches in terms of makespan, energy consumption, and cost. Finally, Section VI concludes the study, summarizing key findings and outlining future research directions.

II. Literature Survey

Many studies have examined methodologies for time and cost-efficient workflow allocation in a cloud computing environment, with limited focus on energy. The hybrid strategies that integrate heuristics like HEFT with metaheuristics have shown promising results in minimizing makespan, energy consumption, and cost simultaneously, forming a strong motivation for the hybrid HEFT-IGWO approach proposed in this study.

Li H et al. [7] proposed the SEPSO (Swarm Entropy-based Particle Swarm Optimization) framework. It is designed to optimize cloud provider profits by minimizing the monetary cost of energy and leased public cloud resources. SEPSO monitors swarm diversity during iterations to avoid premature convergence and enhance exploration by dynamically adjusting search parameters to balance exploration and exploitation throughout the optimization process. Although it improves task sequencing and allocation across public and private clouds, SEPSO mainly targets cost reduction from the provider's view and lacks adaptability to dynamic or large-scale workflows.

Li H et al. [8] proposed the Chaotic-nondominated-sorting Owl Search Algorithm (COSA) to schedule resource-constrained multiple workflows. It combines an Owl Search Algorithm (OSA) with an NSGA-II to optimize makespan, cost, and energy under the specified deadline and budget constraints in hybrid clouds. For better balancing of exploration and exploitation, multiple workflows are scheduled in DVFS-enabled cloud using a hierarchical evolving mechanism that employs COSA for the Worse Half of Population (WHP) and NSGA-II for updating the Better Half of Population (BHP). However, its reliance on fixed workflow and resource settings limits adaptability in dynamic cloud environments, and chaotic control increases computational overhead.

Thekkepuryil et al. [9] proposed a hybrid algorithm combining ALO and PSO to enhance workflow scheduling to optimize key metrics like execution time, cost, and load balancing in cloud environments. The Data Encryption Standard (DES) ensures data security during scheduling, addressing both performance and confidentiality concerns. The hybrid algorithm combines the global search ability of ALO with the local search refinement of PSO. However, its reliance on static workflows and fixed resources limits adaptability in real-world dynamic cloud environments, and the inclusion of DES adds computational overhead.

Saeedi et al. [10] optimizes task-resource mapping in workflows through the following steps. Initialize particles representing task-resource mappings, along with their positions and velocities, evaluate each particle, store non-dominated solutions generated during the search process in external archive, iteratively updates the archive, ideal point, and hyperplane. The crowding distance determines which solutions remain in the archive when it is full. The archive of optimized solutions is returned. However, I-MaOPSO reliance on roulette wheel leader selection can lead to inaccurate results when individuals have similar fitness values, reducing diversity and convergence stability.

Alaei et al. [11] proposed an adaptive fault detection method in cloud computing based on the Improved Differential Evolution (IDE) algorithm combined with an ANFIS prediction model. This hybrid model enhances reliability by predicting faults before they occur, helping reduce makespan, cost, and energy consumption while improving fault tolerance. However, the model primarily addresses VM faults and overlooks other significant reliability issues, such as network, storage, and I/O failures.

Mohammadzadeh et al. [12] proposed the HGALO-SCA algorithm, which addresses the limitations of ALO and SCA algorithms by leveraging their strengths. HGALO-SCA enhances ALO searchability by incorporating SCA oscillatory behaviour, allowing global exploration. The integration of the ALO elite strategy accelerates the convergence while maintaining global search accuracy. The comparison with SPEA2 shows better performance in balancing trade-offs. However, random chaos parameters may also increase computational complexity and reduce consistency in results.

Yao et al. [13] proposed Endocrine-based Coevolutionary Multi-Swarm for Multi-Objective Optimization (ECMSMOO) addresses the workflow scheduling in cloud computing, which is an NP-complete problem by optimizing multiple conflicting objectives, including execution time, cost, and energy consumption. The manager server collects available cloud resources to reduce the impact of elastic resource fluctuations during scheduling. The endocrine-based evolutionary strategy mimics hormone-regulated particle behaviour to enhance search efficiency and avoid local optima. The swarms collaborate and compete to ensure a strong search process across the multi-objective optimization space. Furthermore, the hormonal control mechanism adds algorithmic complexity and overhead, which can hinder real-time scheduling.

Mohammadzadeh et al. [14] introduced an improved version of the GWO algorithm, known as HCGWO. This improvement achieves better optimization outcomes by combining chaos theory and the hill-climbing method. The study extends the proposed IGWO algorithm to a binary version specifically designed for the workflow allocation problem. This involves utilizing various S and V functions to deal with the workflow allocation problem with the goal of reducing execution costs, makespan, and power consumption. chaotic map generation introduces additional computational overhead.

Hassan et al. [15] proposed Smart Energy and Reliability Aware Scheduling (SERAS) algorithm addresses the dual challenge of energy efficiency and system reliability in cloud computing environments. By integrating the DVFS technique, SERAS dynamically adjusts the processor frequencies of VMs while ensuring tasks meet their deadlines. The SERAS divides the deadline across tasks, enabling more precise scheduling and resource allocation. However, algorithm complexity ($O(n^2)$) limits scalability for large workflows.

Belgacem A, Beghdad-Bey K [16] suggested HEFT-ACO, which effectively reduces costs and makespan by combining the ACO algorithm with the HEFT heuristic. The simulation results were conducted on the Amazon EC2 cloud platform using three real-world scientific workflows, viz. Montage, CyberShake, and Ligo show better trade-offs between makespan and cost by leveraging the strengths of both techniques. The workflow characteristics, viz., balance or asymmetry, are not considered.

Taghinezhad A et al. [17] introduce BDCE and BDD, two energy-efficient heuristic algorithms for workflow scheduling in cloud environments, aiming to minimize energy consumption while meeting budget and deadline constraints, target resources with DVFS capabilities, while BDCE works with non-DVFS-enabled resources to optimize cost, scheduling length, and energy savings while ensuring QoS compliance. The result metrics are compared against established methods like BDSD, DBCS, BDHEFT, ERES, and the Safari algorithm using scientific workflow applications, showing improvements in energy savings, cost efficiency, and scheduling performance. However, both BDCE and BDD focus primarily on medium-scale workflows.

Khaleel M [18] proposed the ELSCiW (Energy-Latency-aware Scientific Workflow scheduling) framework focuses on optimizing cloud workflow scheduling by balancing energy consumption and latency while maintaining high QoS. The approach involves a two-step optimization process: first, node efficiency evaluation, which compares the number of handled transactions to power consumption, optimizing processor utilization, and then task mapping with mean GWO, which applies to map tasks to cloud resources efficiently. However, it primarily focuses on the energy–latency trade-off.

Zeedan et al.[19] proposed Enhanced Binary Artificial Bee Colony-based Pareto Front (EBABC-PF) to optimize makespan, processing cost, and resource utilization without violating SLA. The algorithm uses HEFT to prioritize the work, Greedy Randomized Adaptive Search Procedure (GRASP) to develop an initial solution, and Enhanced Binary Artificial Bee Colony (BABC) to schedule tasks onto VMs. Moreover, it emphasizes performance and cost but gives limited attention to energy efficiency.

III. Problem Definition

When a user submits a workflow application as a Directed Acyclic Graph (DAG), as shown in Fig. 1, the workflow interface converts the DAG into an admission queue that maintains execution order based on dependencies. The workflow scheduler allocates tasks to VMs using a scheduling algorithm, optimization criteria for energy efficiency, and a resource estimator to match tasks with suitable resources.

A. Resource Model

In our cloud computing model, the infrastructure comprises a collection of cloud servers (CS), each hosting multiple computing nodes on which VMs are deployed. A centralized global cloud manager oversees the operation, consisting of a resource manager and a scheduler. The resource manager maintains real-time status and key attributes of all available VMs, such as processing speed, energy consumption, and reliability. This information is utilized by the scheduler to efficiently assign tasks when a workflow application is submitted, based on task dependencies and VM availability.

All VMs are assumed to support Dynamic Voltage Scaling (DVS), allowing them to adjust their voltage and frequency levels dynamically. This enables energy savings by switching to the lowest voltage levels during idle periods. The energy consumption model is based on the CMOS power dissipation model, where energy usage is influenced by both the characteristics of the device and the voltage supply associated with each task.

The energy consumption of tasks on VMs during active (busy) execution time and when the VMs are idle can be calculated using:

$$E_{busy} = \sum_{j=1}^{M} \alpha * V_j^2 * ET(T_i, VM_j)$$

$$E_{idle} = \sum_{j=1}^{M} \alpha * V_{j,low}^2 * \Delta idle_j$$

where, n is the total number of tasks, VM_j is the virtual machine on which task $T_i$is executed, V_j is the supply voltage of VM_j, and ET (T_i, VM_j) is the execution time of task T_i on VM_j.$VM_{j,low}$is the minimum voltage level on VM_j, and $\Delta idle_j$ is the amount of idling time for VM_j. The total energy consumption can be calculated as:

$$\text{TEC}=E_{busy}+ E_{idle}$$

B. Application Model

A DAG represents a workflow application represented by Wf = (T, E), where $T = \{T_i, 1 \le i \le \text{N}\}$, is a set of tasks of the workflow, and E is the set of edge characterizes precedence constraints between tasks. The edge $T_i \rightarrow T_j$,shows the precedence relation between T_i and T_j in the DAG as shown in Fig. 2, which will be processed on virtual machines $V = \{VM_j \mid 1 \le j \le \text{M}\}$ with various computing speed and cost. The execution time of task T_i on VM_j is denoted by ET (T_i, VM_j and CT (T_i, T_k) denotes communication. According to the dependencies among the tasks, task T_i (child task) will execute after $Pred(T_i)$, (parent tasks of T_i). The start time ST (T_i, VM_j), finish time $FT(T_i, VM_j)$ of a task T_i on VM_j, and the makespan(MS) are calculated as follows:

$$ST(T_i, VM_j) = max\ \{Avl(VM_j),\ max\{FT[Pred(T_i)] + CT(Pred(T_i), T_k)\}$$

$$FT(T_i, VM_j) = ST(T_i, VM_j) + ET(Pred(T_i), VM_j)$$

$$Makespan\ (MS) := max \sum_{i=1}^{N} FT(T_i, VM_j)$$

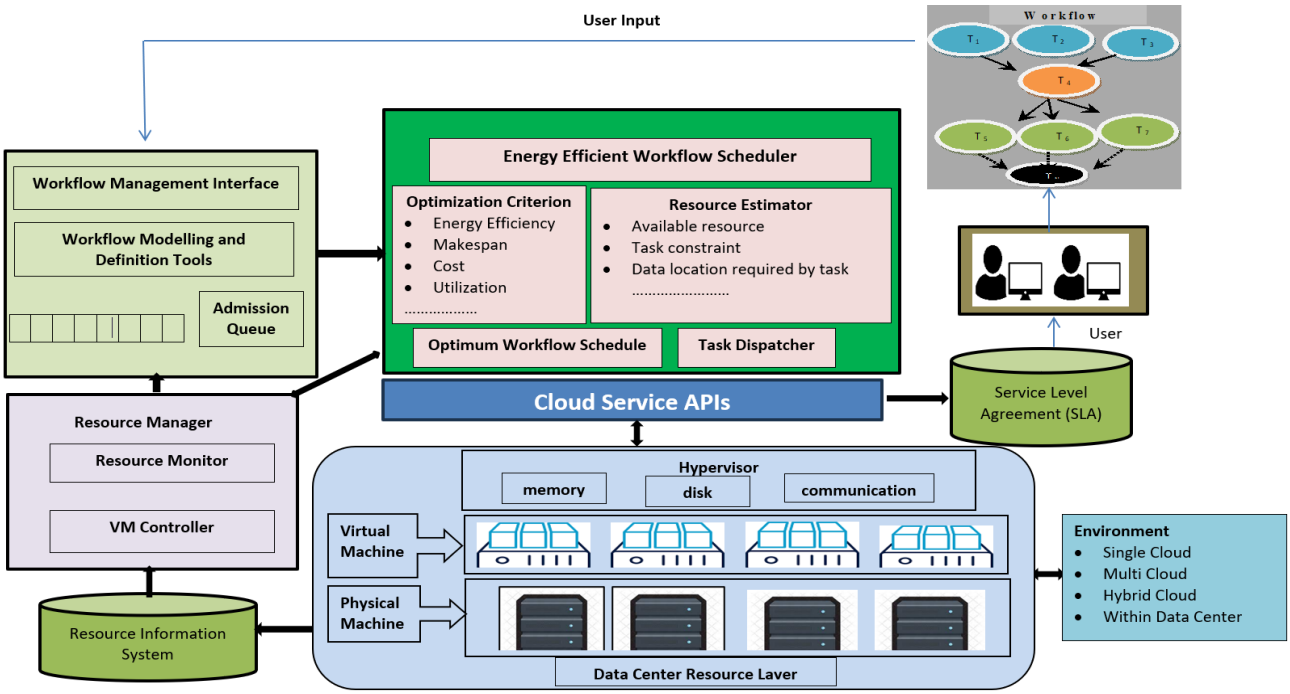


Fig. 1. Workflow allocation framework.

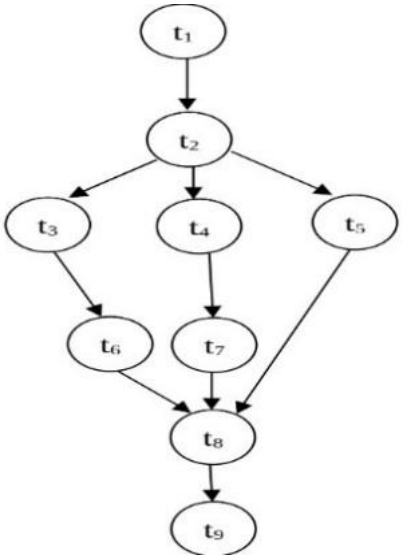


Fig. 2. Sample workflow.

C. Cost Model

The pricing in cloud computing is determined by the billing interval (BI) for resource leasing. As a result, even if only a portion of the time is used, the client must pay for the full interval.

$$Cost(T_i, VM_j) = \sigma * \left\lceil \frac{ET\ (T_i, VM_j)}{\tau} \right\rceil * P_0 * e^{\frac{cpu\ cycles\ of\ VM_j}{slowest\ cpu\ cycle}}$$

where, σ is a random variable used to generate different combinations of VM pricing and capacity, τ unit chargeable time, P_0 is the base price of VM. The total execution cost can be defined as:

Let Bi, j be a Boolean variable, such that

$$Bi,j = \begin{cases} 1, & if\ task\ T_i\ is\ assign\ to\ V_j \\ 0, & otherwise \end{cases}$$

$$TCost = \sum_{i=1}^{N} \sum_{j=1}^{M} Bi,j * Cost(T_i, VM_j)$$

D. Problem Definition

A schedule (Sc) is presented as Sc = (*SVM, Allocation, MS*, $TCost$, TEC), where *SVM* corresponds to virtual machines, *Allocation* represents the task to VM mappings, $TCost$: Total execution cost, *MS*: Makespan, and TEC: Energy consumption of the workflow. The problem can be defined as:

Minimize:

$$F = \gamma \text{ x } MS + \vartheta \text{ x } TCost + (1 - \gamma - \vartheta) \text{ x } TEC$$

Subject to

$$i)\ \sum_{j=1}^{M} Bi,j = 1,\ i = 1,2,3.....n$$

$$ii)\ 0 < \gamma, \vartheta < 1$$

Constraints: i) indicate that any task can be assigned to only one VM and the constraint, and ii) limit the range of γ, ϑ that balances optimization functions.

IV. Algorithm

The Grey Wolf Optimization algorithm is a meta-heuristic optimization technique that was inspired by the social hierarchy and hunting habits of grey wolves. It was first presented by Mirjalili et al. in 2014 [20]. We propose an energy-efficient workflow allocation algorithm based on improving the Grey Wolf Optimizer is a version of GWO that has been created to improve its search performance. This algorithm has gained significant attention due to its simplicity, flexibility, and

effectiveness in solving various complex optimization problems. GWO mimics the leadership structure and collaborative hunting strategy of grey wolves, which enhances its ability to explore and exploit the search space efficiently. In a grey wolf pack, there are four types of wolves, as shown in Fig. 3, based on their social hierarchy:

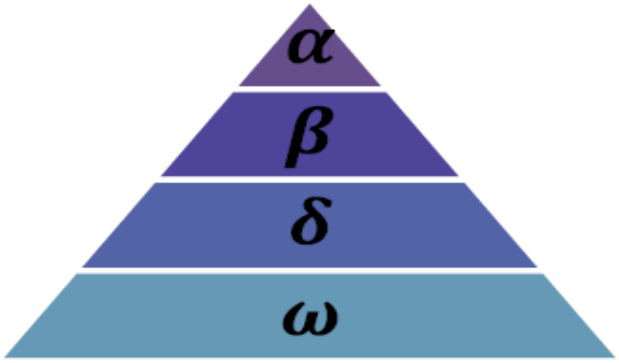


Fig. 3. Hierarchy flow up for the GWO algorithm.

Alpha (α): The leader of the pack, responsible for making decisions and guiding the hunting process.

Beta (β): The second-in-command, assisting the alpha in decision-making and maintaining discipline in the pack.

Delta (δ): The third level in the hierarchy, responsible for reporting to the alpha and beta and assisting them in their tasks.

Omega (ω): The lowest-ranking wolves, following the orders of the higher-ranking wolves and playing a crucial role in the pack's social structure.

The following mathematical models of GWO are useful in different scenarios of any real-world application.

1) Encircling prey: Grey wolves encircle their prey during the hunt. This encircling behaviour is mathematically modelled as follows:

$$\vec{D} = |\vec{C} * \overrightarrow{X_p}(t) - \vec{X}(t1)|$$

$$\vec{X}(t+1) = \overrightarrow{X_p}(t) - \vec{A} * \vec{D}$$

Here, t represents the current iteration, $\overrightarrow{X_p}$ is the position vector of the prey, $\vec{X}$ is the position vector of a grey wolf, and $\vec{A}$ and $\vec{C}$ are coefficient vectors calculated as:

$$\vec{A} = 2\,\vec{a} * \overrightarrow{r_1} - \vec{a}$$

$$\vec{C} = 2\,\overrightarrow{r_2}$$

$$\mathrm{a} = 2\left(1 - \frac{t}{T}\right)$$

In these equations, $\vec{a}$ linearly decreases from 2 to 0 over the course of iterations, while $\overrightarrow{r_1}$ and $\overrightarrow{r_2}$ are random vectors within the range [0, 1].

2) Hunting: The hunting phase involves following the best individuals in the pack, specifically the α, β, and δ wolves. The positions are updated according to these leaders as:

$$\overrightarrow{D_\alpha} = |\overrightarrow{C_1} * \overrightarrow{X_\alpha} - \vec{X}|$$

$$\overrightarrow{D_\beta} = |\overrightarrow{C_2} * \overrightarrow{X_\beta} - \vec{X}|$$

$$\overrightarrow{D_\delta} = |\overrightarrow{C_3} * \overrightarrow{X_\delta} - \vec{X}|$$

$$\overrightarrow{X_1} = \overrightarrow{X_\alpha} - \overrightarrow{A_1} * \overrightarrow{D_\alpha}$$

$$\overrightarrow{X_2} = \overrightarrow{X_\beta} - \overrightarrow{A_2} * \overrightarrow{D_\beta}$$

$$\overrightarrow{X_3} = \overrightarrow{X_\delta} - \overrightarrow{A_3} * \overrightarrow{D_\delta}$$

The position of a grey wolf is then updated as follows:

$$\vec{X}(t+1) = \frac{\overrightarrow{X_1} + \overrightarrow{X_2} + \overrightarrow{X_3}}{3}$$

3) Attacking prey: The attacking phase is indicated by reducing the value of $\vec{a}$ as iterations progress. When $\vec{A}$ falls within the range [-1, 1], wolves are encouraged to attack the prey by shrinking their encircling behaviour.

Search for Prey: When $|\vec{A}|>1|$, wolves diverge to explore new areas.

The improved GWO using differential evolution for stronger exploration and diversification through mutation, crossover, and survival of the fittest for maintaining the best-performing solutions and eliminating poor-quality candidates. Additionally, we introduce a hybrid initialization technique by incorporating a solution generated using the HEFT algorithm, ensuring that the search starts from at least one good-quality position. The population of N wolves (solutions) is initialized randomly. Each wolf represents a mapping of tasks to VMs. To enhance quality and convergence, one individual is replaced by a HEFT-based solution, leveraging domain knowledge to start the search from a near-optimal region. Each solution is evaluated using a fitness function based on makespan, cost, and energy. Then sort the wolves based on their fitness values and choose the best $\overrightarrow{X_\alpha}$, second best $\overrightarrow{X_\beta}$, and third best $\overrightarrow{X_\delta}$. During each iteration, the positions of the wolves are updated using the DE strategy. This involves mutating the positions to create new candidates, crossing over elements between pairs of candidates to combine their features, and selecting the best candidates to form the next generation.

4) Mutation operation: The mutation operation of differential evolution is its most significant feature. To generate variance in this process, two weighted difference vectors are appended to the selected individual. The difference vector from the parents, which consists of two different individuals $(X_{r_1}^t, X_{r_2}^t)$ from the parent generation (the t-th generation), is the basic component of the DE variation mechanism. The definition of the difference vector is as follows:

$$Dd_{r_{12}} = X_{r_1}^t - X_{r_2}^t$$

where, r_1, r_2 are index numbers of distinct population members, and as a result, the mutation operation can be written as follows:

$$V_i^{t+1} = X_{r_3}^t + \mathrm{F} * (X_{r_1}^t - X_{r_2}^t)$$

where, F is the scaling factor, and r_1, r_2 and r_3 are different integers in the scope $(1, 2, \ldots n)$ from the current target vector index(i).

5) Crossover operation: A crossover operation is carried out using the mutation vector X_i^{t+1} for the individual X_i^t wolves, producing a trial individual U_i^{t+1}. A random selection technique is used to make sure that at least one bit of U_i^{t+1} is derived from V_i^{t+1} to ensure that X_i^t evolves. The crossover probability factor (CR) for the remaining bits of U_i^{t+1} determines whether each bit originates from X_i^t or V_i^{t+1}. The expression for the crossover operation as follows:

$$U_{ij}^{t+1} = \begin{cases} V_{ij}^{t+1}, rand(\mathrm{j}) \leq CR \ OR \ J = randn(i) \\ X_{ij} \quad , rand(\mathrm{j}) > CR \ AND \ J \neq randn(i) \end{cases}$$

$$J = 1, 2, .., \mathrm{D}$$

where, rand(j) ∈ [0, 1] obeys the random-uniform distribution, j is the j-th variable (gene), CR is the crossover probability, and rand(i) ∈ [1, 2, … D].

6) Selection operation: The greedy choice used in the selection process. Following the mutation and crossover operations, the trial individual U_i^{t+1} is created and then compared with the target individual X_i^t. The comparison can be expressed numerically as:

$$X_i^{t+1} = \begin{cases} U_i^{t+1}, & f(U_i^{t+1}) < f(X_i^t) \\ X_i^t & f(U_i^{t+1}) \geq f(X_i^t) \end{cases} \quad i = 1,2, \ldots n$$

In summary, the proposed enhances the original GWO by incorporating evolutionary principles and the DE strategy, significantly improving its search performance. This improved algorithm effectively balances exploration and exploitation, making it a powerful tool for solving complex optimization problems.

An illustration to explain HEFT-IGWO, traditional GWO, and HEFT has been presented in this section with a random workflow consisting of nine tasks with precedence among themselves, as shown in Fig. 2. We have taken only three VMs, each with different computing capacities, energy consumption, and cost, but there may be more in real scenarios. Each wolf in the GWO represents a possible allocation of tasks to VMs.

V. Results and Discussion

We use simulation for the performance study of proposed algorithms with the traditional GWO algorithm and HEFT algorithms. The simulation was performed using Python on a system with an i5 processor having a frequency of 2.1 GHz and 16 GB RAM running Windows 11. A random workflow application used to assess the HEFT-IGWO algorithm. Table I shows the simulation parameters and their corresponding values.

TABLE I Simulation Parameters

Parameter	Values
Tasks (N) number	5-100
VMs (M) number	3
Processing Speed	1000-5000 MIPS
VM distances	1-100
Computing Capacity of VM	30-1000
Inter task communication	1-1000

This section provides an overview of the performance findings in terms of makespan, energy, and cost, corresponding to three virtual machines. The performance of the three algorithms was evaluated in terms of makespan for different numbers of tasks ranging from 5 to 100 is shown in Fig. 4. The HEFT algorithm serves as the baseline for comparison.

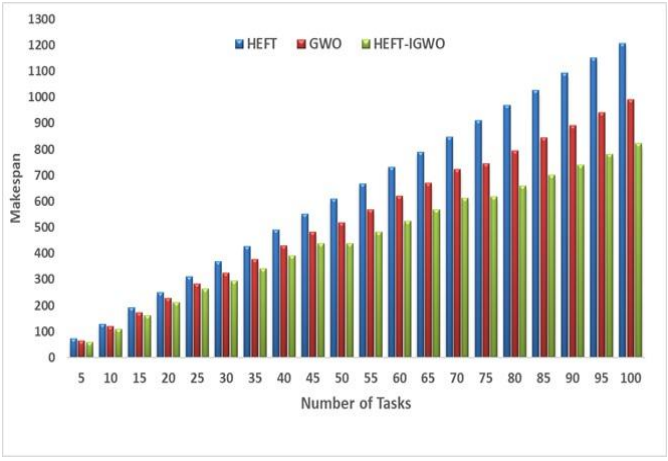


Fig. 4. Makespan vs. Number of workflow tasks.

Its makespan values increase almost linearly with the number of tasks. For small workflows (5 to 25 tasks), HEFT performs reasonably well because the search space is limited, and its ranking and earliest finish time heuristic is efficient. However, as the number of workflow tasks increases (50 to 100), HEFT performance degrades significantly, producing the highest makespan values among the three methods. For small to medium workflows (5 to 45 tasks), GWO provides an average improvement of around 8–12% in makespan. As the number of workflow tasks increases, GWO advantage becomes more significant due to its iterative exploration and leader based hunting mechanism, which helps find more balanced task–VM mappings. For larger workloads (50 to 100 tasks), the improvement rises to 15–18%. The Hybrid HEFT–IGWO approach achieves the best performance across all workflows, demonstrating the benefit of combining HEFT based initialization with improved GWO exploration through differential evolution operators and survival of the fittest. For small and medium workflows (5 to 45 tasks), HEFT–IGWO improves makespan by 15–20% compared to HEFT. For large workflows (50 to 100 tasks), the improvements become more substantial, ranging from 28–32%. These significant gains are due to the ability to start with a near optimal HEFT solution and then further refine allocations using DE mutation, crossover, and iterative leader updates.

ALGORITHM 1: HEFT-IGWO

Input: Workflow W (T, E), VM set V, population size N, generations G
Output: Best schedule with minimized makespan, cost, and energy
Phase 1: Initialization
Compute upward ranks of tasks using HEFT ordering.
Seed HEFT-based elite solution as one candidate wolf.
Generate N-1 random task-to-VM mappings as additional wolves.
Evaluate fitness of all wolves (makespan, energy, cost).
Identify alpha, beta, delta leaders (best three wolves).

Phase 2: Iterative Hybrid Evolution
For gen = 1 To G do
Update wolves positions using GWO encircling and hunting rules.
Apply higher probability mutation (random task VM flips).
Perform DE style discrete crossover between wolves and mutants.
Evaluate fitness of offspring (makespan, cost, energy).
Apply survival-of-the-fittest: keep best between parent and offspring.
Update leaders alpha, beta, delta based on new population.
Log convergence metrics (fitness, makespan, energy, cost).

Phase 3: Termination
Return alpha wolf (best schedule) with final metrics.
end

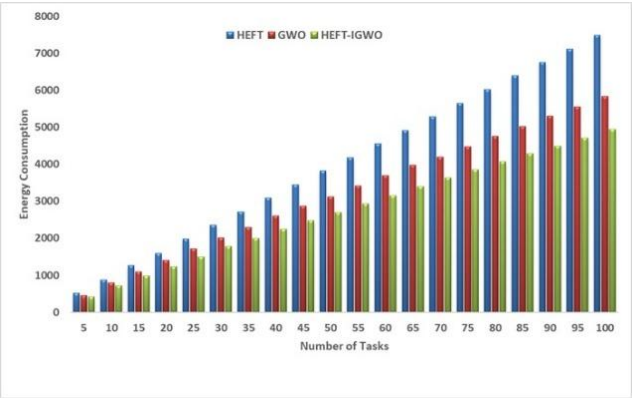


Fig. 5. Energy consumption vs. Number of workflow tasks.

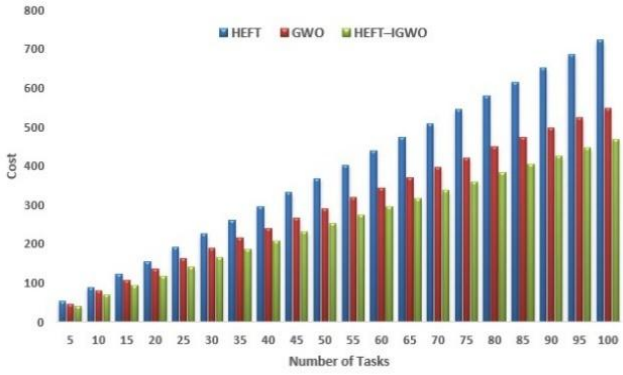


Fig. 6. Cost vs. Number of workflow tasks.

All algorithms show increasing energy consumption as tasks increase, but the growth rate is significantly slower for GWO and HEFT–IGWO than HEFT as shown in Fig. 5. The HEFT algorithm consistently consumes the highest amount of energy. The GWO achieves clear energy reductions across all task sizes, averaging around 10–15% for smaller tasks (5–30) and 18–22% for larger workloads (50–100). The hybrid HEFT–IGWO algorithm (see Algorithm 1) consistently outperforms both HEFT and GWO, demonstrating the largest energy savings across all workflow sizes. For small and medium workflow (5–40 tasks), energy savings are around 20–27%, while for large workloads (50–100 tasks), savings increase to 30–34%. HEFT–IGWO achieves the largest and most consistent energy reductions, particularly for large scale workflows, making it ideal for energy-aware workflow allocation in cloud environments.

Since HEFT focuses on earliest finish time without explicitly optimizing cost, it often results in inefficient VM usage leading to higher costs as workflows scale as shown in Fig. 6. The GWO consistently reduces operating cost compared to HEFT. For the small workflows (5–25 tasks), the cost improvements are in the range of 10–14%. For medium workloads (30–60 tasks), cost improvements increase to around 17–22%, and for larger workloads (65–100 tasks), savings stabilize between 22–24%. The hybrid HEFT–IGWO algorithm provides the largest cost reductions across all task sizes. For small tasks, improvements are already noticeable at ~20–22%, increasing to 26–30% for medium workflows, and reaching ~35% for larger workloads.

VI. Conclusion

This study presented a comparative evaluation of three workflow scheduling techniques HEFT, GWO, and the proposed Hybrid HEFT–IGWO to enhance performance and energy efficiency in cloud computing environments. The proposed HEFT–IGWO hybrid delivers the most notable results. By initializing the search with a HEFT based schedule and refining it through differential evolution and survival-of-the-fittest operations, it effectively balances exploration and exploitation. The hybrid model achieves up to 32% reduction in makespan, 34% lower energy consumption, and 35% cost savings, outperforming both baseline methods across all task scales. Overall, combining heuristic and evolutionary intelligence significantly enhances workflow allocation efficiency. HEFT–IGWO inherits the computational speed of HEFT while leveraging IGWO adaptive global search to achieve optimized, energy aware scheduling. Future work will focus on extending this approach to dynamic, heterogeneous, and QoS driven cloud environments for real time workflow.

References

[1] B. Furht and A. Escalante, Eds., Handbook of Cloud Computing. Boston, MA: Springer US, 2010. doi: 10.1007/978-1-4419-6524-0.

[2] M. Alam, S. Mustajab, M. Shahid, and F. Ahmad, "Cloud Computing: Architecture, Vision, Challenges, Opportunities, and Emerging Trends," in 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), IEEE, Nov. 2023, pp. 829–834. doi: 10.1109/ICCCIS60361.2023.10425507.

[3] M. M. Nezami, A. Kumar, M. Shahid, and M. M. Nezami, "Analysis of Energy Efficient Workflow Allocation in Cloud Computing," in 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), IEEE, Apr. 2023, pp. 803–809. doi: 10.1109/CSNT57126.2023.10134644.

[4] P. A. Malla, S. Sheikh, M. Shahid, and S. U. Mushtaq, "Energy-efficient sender-initiated threshold-based load balancing e-STLB in cloud computing environment," Concurr Comput, vol. 36, no. 5, Feb. 2024, doi: 10.1002/cpe.7943.

[5] M. Sajid and Z. Raza, "Energy-aware stochastic scheduling model with precedence constraints on DVFS-enabled processors," Turkish Journal of Electrical Engineering and Computer Sciences, vol. 24, no. 5, pp. 4117–4128, 2016, doi: 10.3906/elk-1505-112.

[6] K. K. Chakravarthi, P. Neelakantan, L. Shyamala, and V. Vaidehi, "Reliable budget aware workflow scheduling strategy on multi-cloud environment," Cluster Comput, vol. 25, no. 2, pp. 1189–1205, Apr. 2022, doi: 10.1007/S10586-021-03464-4.

[7] H. Li, X. Li, J. Xu, and L. Chen, "Entropy based swarm intelligent searching for scheduling deadline constrained workflows in hybrid cloud," International Journal of Machine Learning and Cybernetics, vol. 15, no. 4, pp. 1183–1199, Apr. 2024, doi: 10.1007/s13042-023-01962-y.

[8] H. Li, G. Xu, D. Wang, M. Zhou, Y. Yuan, and A. Alabdulwahab, "Chaotic-Nondominated-Sorting Owl Search Algorithm for Energy-Aware Multi-Workflow Scheduling in Hybrid Clouds," IEEE Transactions on Sustainable Computing, vol. 7, no. 3, pp. 595–608, Jul. 2022, doi: 10.1109/TSUSC.2022.3144357.

[9] J. Kakkottakath Valappil Thekkepuryil, D. P. Suseelan, and P. M. Keerikkattil, "An effective meta-heuristic based multi-objective hybrid optimization method for workflow scheduling in cloud computing environment," Cluster Comput, vol. 24, no. 3, pp. 2367–2384, 2021, doi: 10.1007/s10586-021-03269-5.

[10] S. Saeedi, R. Khorsand, S. Ghandi Bidgoli, and M. Ramezanpour, "Improved many-objective particle swarm optimization algorithm for scientific workflow scheduling in cloud computing," Comput Ind Eng, vol. 147, p. 106649, Sep. 2020, doi: 10.1016/j.cie.2020.106649.

[11] M. Alaei, R. Khorsand, and M. Ramezanpour, "An adaptive fault detector strategy for scientific workflow scheduling based on improved differential evolution algorithm in cloud," Appl Soft Comput, vol. 99, p. 106895, Feb. 2021, doi: 10.1016/J.ASOC.2020.106895.

[12] A. Mohammadzadeh, M. Masdari, F. S. Gharehchopogh, and A. Jafarian, "A hybrid multi-objective metaheuristic optimization algorithm for scientific workflow scheduling," Cluster Comput, vol. 24, no. 2, pp. 1479–1503, Jun. 2021, doi: 10.1007/S10586-020-03205-Z.

[13] G. Yao, Y. Ding, Y. Jin, and K. Hao, "Endocrine-based coevolutionary multi-swarm for multi-objective workflow scheduling in a cloud system," Soft Computing - A Fusion of Foundations, Methodologies and Applications, vol. 21, no. 15, pp. 4309–4322, Aug. 2017, doi: 10.1007/S00500-016-2063-8.

[14] A. Mohammadzadeh, M. Masdari, F. S. Gharehchopogh, and A. Jafarian, "Improved chaotic binary grey wolf optimization algorithm for workflow scheduling in green cloud computing," Evol Intell, vol. 14, no. 4, pp. 1997–2025, Dec. 2021, doi: 10.1007/s12065-020-00479-5.

[15] H. A. Hassan, S. A. Salem, and E. M. Saad, "A smart energy and reliability aware scheduling algorithm for workflow execution in DVFS-enabled cloud environment," Future Generation Computer Systems, vol. 112, pp. 431–448, Nov. 2020, doi: 10.1016/j.future.2020.05.040.

[16] A. Belgacem and K. Beghdad-Bey, "Multi-objective workflow scheduling in cloud computing: trade-off between makespan and cost," Cluster Comput, vol. 25, no. 1, pp. 579–595, Feb. 2022, doi: 10.1007/s10586-021-03432-y.

[17] A. Taghinezhad-Niar, S. Pashazadeh, and J. Taheri, "Energy-efficient workflow scheduling with budget-deadline constraints for cloud," Computing, vol. 104, no. 3, pp. 601–625, Mar. 2022, doi: 10.1007/s00607-021-01030-9.

[18] M. I. Khaleel, M. Safran, S. Alfarhood, and M. Zhu, "Energy-latency trade-off analysis for scientific workflow in cloud environments: The role of processor utilization ratio and mean grey wolf optimizer," Engineering Science and Technology, an International Journal, vol. 50, p. 101611, Feb. 2024, doi: 10.1016/j.jestch.2023.101611.

[19] M. Zeedan, G. Attiya, and N. El-Fishawy, "Enhanced hybrid multi-objective workflow scheduling approach based artificial bee colony in cloud computing," Computing, 2022, doi: 10.1007/S00607-022-01116-Y/FULLTEXT.HTML.

[20] 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.

Feature Pyramid Network with Dual-Decoder Supervision for Accurate Stroke Lesion Localization in Multi-Modal Brain MRI

Satmyrza Mamikov[1], Zhansaya Yakhiya[2], Bauyrzhan Omarov*[3],
Yernar Mamashov[4], Akbayan Aliyeva[5], Balzhan Tursynbek[6]
University of Friendship of People's Academician A. Kuatbekov, Shymkent, Kazakhstan[1, 2]
Al-Farabi Kazakh National University, Almaty, Kazakhstan[3, 4]
Narxoz University, Almaty, Kazakhstan[3]
Khoja Akhmet Yassawi International Kazakh-Turkish University, Turkistan, Kazakhstan[5, 6]

***Abstract*—This study presents a novel Feature Pyramid Network with Dual-Decoder Supervision for accurate stroke lesion localization in multi-modal brain MRI. The proposed architecture integrates a Swin Transformer backbone with multi-scale feature aggregation, enabling effective fusion of hierarchical representations from DWI, ADC, and FLAIR sequences. A dual-decoder structure is employed, where the auxiliary decoder provides coarse lesion guidance through pseudo masks, and the primary decoder refines boundaries for precise voxel-level segmentation. Auxiliary supervision improves convergence stability and feature discrimination, while modality dropout enhances robustness to incomplete imaging protocols. Experiments conducted on the ATLAS v2.0 dataset demonstrate superior performance over baseline encoder–decoder models, achieving higher Dice scores, improved boundary accuracy, and strong lesion-wise detection rates. The model consistently localizes lesions of varying size, shape, and intensity, with minimal overfitting, as evidenced by small training–testing performance gaps. Qualitative results confirm the framework's ability to transform coarse localization into anatomically accurate predictions. The combination of multi-modal integration, dual-decoder specialization, and self-training mechanisms positions the proposed method as a promising candidate for clinical deployment in rapid stroke diagnosis workflows. Future directions include expanding validation to multi-center datasets, incorporating explainable AI techniques, and enabling real-time 3D processing for deployment in acute care environments.**

***Keywords*—Stroke lesion localization; multi-modal MRI; feature pyramid network; segmentation; deep learning**

I. Introduction

Stroke remains one of the leading causes of mortality and long-term disability worldwide, with timely and precise lesion localization being a critical determinant for successful therapeutic intervention [1]. Magnetic Resonance Imaging (MRI) is the preferred non-invasive modality for stroke diagnosis due to its superior soft-tissue contrast and ability to capture diverse tissue characteristics across multiple imaging sequences [2]. Accurate lesion segmentation and localization not only assist in diagnosis but also facilitate the evaluation of stroke severity, prognosis prediction, and treatment planning [3]. However, automated stroke lesion detection poses substantial challenges owing to the heterogeneity of lesion shapes, sizes, and intensities, as well as the presence of noise, motion artifacts, and variations across different MRI modalities [4]. Addressing these challenges requires models that can robustly integrate multi-scale contextual features while maintaining fine-grained spatial resolution.

Deep learning techniques, particularly convolutional neural networks (CNNs), have shown remarkable success in various medical image analysis tasks, including tumor segmentation, organ delineation, and lesion detection [5]. Yet, traditional encoder–decoder CNN architectures often struggle to capture both global semantic context and detailed boundary information when dealing with complex and irregularly shaped stroke lesions [6]. Feature Pyramid Networks (FPNs) have emerged as a powerful architectural design to mitigate this limitation by enabling multi-scale feature fusion, thereby improving detection and segmentation performance across varying lesion sizes [7]. Despite these advancements, single-decoder frameworks can underutilize the rich hierarchical features extracted by the backbone, leading to suboptimal boundary refinement and reduced robustness in heterogeneous imaging conditions [8].

Recent studies have explored multi-head or multi-decoder architectures to enhance learning by incorporating specialized branches for distinct but complementary tasks, such as coarse lesion localization and fine-grained segmentation [9]. The dual-decoder paradigm facilitates task-specific feature optimization, allowing one branch to focus on high-level semantic structure while the other emphasizes spatial detail preservation. When coupled with auxiliary supervision strategies, this approach can guide intermediate layers toward more discriminative feature representations and accelerate convergence during training. Furthermore, integrating such architectures with multi-modal MRI data such as diffusion-weighted imaging (DWI), apparent diffusion coefficient (ADC) maps, and fluid-attenuated inversion recovery (FLAIR) sequences can significantly boost lesion detectability by leveraging the complementary tissue contrast characteristics inherent in different modalities [10]. This multi-modal fusion, however, demands careful architectural design to avoid

information redundancy and overfitting, especially in datasets with limited sample sizes.

II. Related Works

Automated stroke lesion segmentation has been extensively studied in recent years, driven by advances in deep learning and the availability of annotated neuroimaging datasets [10]. Early methods relied on traditional image processing pipelines, integrating intensity thresholding, region growing, and atlas-based priors [11]. While effective for well-defined lesions, these approaches often failed under conditions of low contrast and irregular lesion morphology [12]. The emergence of deep convolutional neural networks (CNNs) introduced the capability to learn hierarchical features directly from data, enabling better generalization to unseen cases [13]. Architectures such as U-Net and its derivatives became popular for medical image segmentation due to their encoder–decoder structure and skip connections [14]. However, their performance still degraded in multi-modal MRI settings without tailored fusion strategies [15].

Multi-modal MRI analysis has gained attention due to the complementary information provided by sequences like DWI, ADC, and FLAIR [16]. Fusion strategies for these modalities range from simple channel concatenation to more sophisticated attention-based feature integration [17]. Studies have shown that modality-specific feature extractors combined with shared decoding networks can significantly improve segmentation performance [18]. Nevertheless, straightforward concatenation can introduce redundancy and lead to overfitting, particularly in small datasets [19]. Attention-based fusion mechanisms have been applied to mitigate this by selectively weighting modality contributions [20]. Despite these advances, the integration of multi-scale feature representations from multi-modal data remains a challenging and less explored problem in stroke lesion localization [21].

The incorporation of Feature Pyramid Networks (FPNs) into medical imaging pipelines has proven effective in capturing multi-scale contextual information [22]. FPNs enable the aggregation of high-resolution spatial features with deep semantic features, improving detection and segmentation tasks across varying object sizes [23]. In stroke imaging, multi-scale architectures help in detecting both small cortical infarcts and larger subcortical lesions [24]. However, many FPN-based designs in medical imaging rely on single decoder pathways, which may underutilize the available hierarchical features [25]. Dual-decoder approaches have been proposed to address this, separating the tasks of coarse localization and precise segmentation [26]. This separation allows each decoder to specialize, but often lacks coordinated supervision, leading to suboptimal synergy between the two outputs [27].

Auxiliary supervision and multi-task learning strategies have emerged as effective means to guide intermediate network layers toward more discriminative feature representations [28]. By introducing additional loss functions at various stages, these methods encourage the network to learn robust features for both global context and local detail preservation. In the context of stroke lesion analysis, auxiliary segmentation branches have been used to stabilize training and improve boundary accuracy [29]. Teacher–student frameworks have also been integrated with auxiliary supervision to leverage pseudo-labels for semi-supervised learning [30]. Despite promising results, there remains a lack of dedicated architectures that combine FPN, dual-decoder design, and auxiliary supervision specifically optimized for multi-modal stroke lesion localization, representing the gap addressed by the present study [31].

III. Methodology

The proposed Feature Pyramid Network with Dual-Decoder Supervision for stroke lesion localization in multi-modal brain MRI is designed to integrate multi-scale feature representations with specialized decoding paths for enhanced segmentation accuracy. The architecture, illustrated in Fig. 1, is built on a Swin Transformer backbone with FPN for hierarchical feature fusion. It consists of a primary decoder for fine-grained segmentation, an auxiliary decoder for coarse lesion supervision, and a classification head for lesion presence verification.

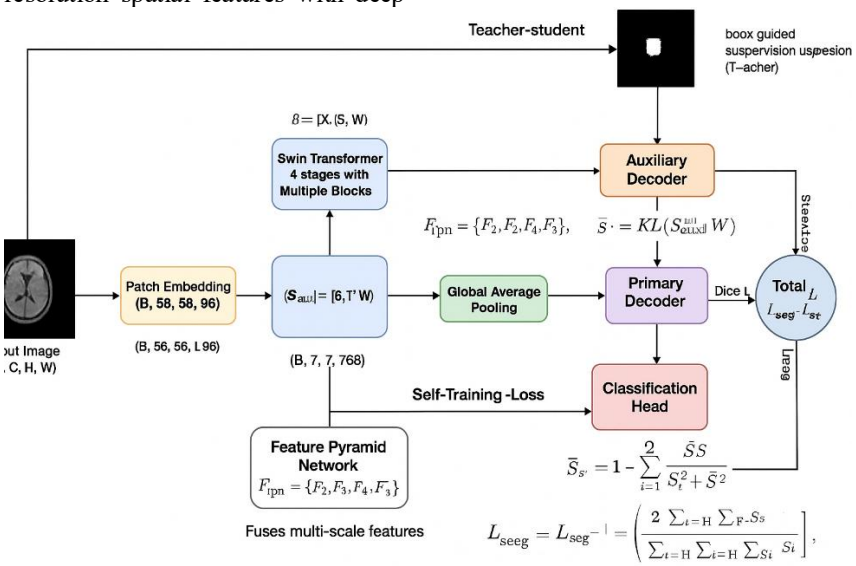


Fig. 1. Example of a figure caption.

A. Input Processing and Patch Embedding

Given a multi-modal MRI volume $X \in R^{C\times H\times W}$ where C denotes number of modalities, the patch embedding module applies a convolution with kernel size k=4 and stride s=4 transforming the image into non-overlapping patches:

$$P = Conv_{k=4,s=4}(X),\ P \in R^{B\times N\times D} \tag{1}$$

where B is the batch size, N is the number of patches, and D is the embedding dimension.

B. Swin Transformer with FPN

The embedded patches are processed through four hierarchical stages of the Swin Transformer, generating feature maps $\{F_2, F_3, F_4, F_5\}$ with progressively reduced spatial resolution and enriched semantic information. The FPN aggregates these:

$$F_{fpn} = FPN(F_2, F_3, F_4, F_5) \tag{1}$$

This yields multi-scale features for both segmentation and classification tasks.

C. Dual-Decoder Structure and Auxiliary Supervision

The primary decoder focuses on boundary-preserving fine segmentation, while the auxiliary decoder provides intermediate supervision for coarse lesion localization. The auxiliary output *S*aux is guided by a pseudo-label mask *W* from a teacher–student model using Kullback–Leibler [32] divergence:

$$L_{aux} = KL(S_{aux} \,||\, W) \tag{2}$$

The final segmentation loss combines Dice and Binary Cross-Entropy (BCE) losses:

$$L_{seg} = L_{Doce} + L_{BCE} \tag{3}$$

With the Dice loss defined as:

$$L_{Dice} = 1 - \frac{2\sum_i p_i g_i}{\sum_i p_i^{\,2} + \sum_i g_i^{\,2}} \tag{4}$$

Where p_i and g_i are predicted and ground-truth voxel probabilities.

D. Classification Head and Self-Training Loss

A classification branch with global average pooling predicts the lesion presence probability. Self-training is incorporated by enforcing consistency between the primary and auxiliary decoders:

$$L_{st} = \left\| S_{primary} - S_{aux} \right\|_2^2 \tag{5}$$

The total loss is:

$$L_{total} = L_{seg} + \lambda_1 L_{aux} + \lambda_2 L_{st} \tag{6}$$

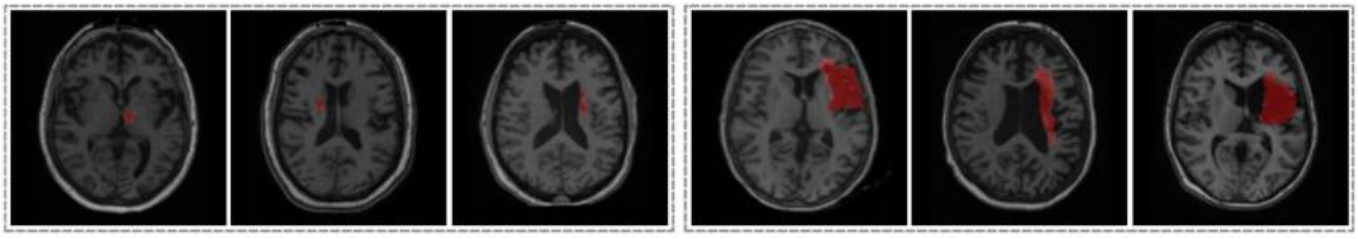
Where λ_1 and λ_2 are balancing weights.

E. Training and Inference Strategy

During training, multi-modal MRI sequences are concatenated channel-wise, with modality-specific augmentations applied to improve generalization. The teacher–student framework updates the teacher via exponential moving average (EMA) of the student weights. At inference, multi-scale test-time augmentation (TTA) is applied, and connected component filtering removes spurious predictions.

F. Dataset

In this study, we employ the ATLAS v2.0 (Anatomical Tracings of Lesions After Stroke) dataset [33], which provides a large collection of clinically acquired structural MRI scans with expert-annotated lesion masks. This dataset contains T1-weighted images from individuals with subacute and chronic stroke, complemented by detailed voxel-level delineations verified through multi-rater consensus. All scans are pre-aligned to MNI-152 space, ensuring anatomical consistency across subjects and facilitating integration into deep learning pipelines. For our task, the T1 modality is utilized in combination with lesion annotations to train and validate the proposed Feature Pyramid Network with dual-decoder supervision. Preprocessing includes N4 bias field correction, skull stripping, intensity normalization to zero mean and unit variance, and resampling to isotropic voxel dimensions to ensure uniformity in spatial resolution.



a) Small lesions b) Big lesions

Fig. 2. Atlas v2.0 dataset samples.

Fig. 2 provides representative samples from the ATLAS v2.0 dataset, demonstrating the variability in lesion size, morphology, and location encountered in real-world stroke presentations. The lesions range from focal cortical damage to extensive subcortical infarctions, capturing the heterogeneity necessary for training robust lesion localization models. This diversity ensures that the proposed Feature Pyramid Network with dual-decoder supervision is tested against a broad

spectrum of stroke manifestations, enabling evaluation of both its fine-grained segmentation capabilities and its adaptability to challenging anatomical contexts. By leveraging these standardized and expertly annotated samples, the dataset supports reproducible, high-quality benchmarking of advanced deep learning architectures for stroke lesion analysis.

IV. Results

The Results section presents a comprehensive evaluation of the proposed Feature Pyramid Network with Dual-Decoder Supervision for stroke lesion localization in multi-modal brain MRI. This section reports both quantitative performance metrics and qualitative visual analyses, enabling a detailed assessment of the model's accuracy, generalization, and robustness. The results are organized to first illustrate the model's convergence behavior during training, followed by segmentation and classification performance compared to baseline and state-of-the-art methods. Additionally, visual examples are provided to demonstrate the network's ability to refine coarse pseudo masks into anatomically precise segmentations and to generate accurate lesion bounding boxes. Together, these findings validate the effectiveness of the proposed approach and highlight its potential applicability in real-world clinical workflows for rapid and reliable stroke diagnosis.

A. Evaluation Parameters

To rigorously assess the performance of the proposed Feature Pyramid Network with Dual-Decoder Supervision for stroke lesion localization in multi-modal brain MRI, several quantitative evaluation metrics are employed. These parameters are selected to measure both voxel-level segmentation quality and lesion-wise detection accuracy, ensuring a comprehensive analysis of the model's performance.

The DSC [33] evaluates the spatial overlap between the predicted lesion mask P and the ground truth G, defined as:

$$DSC = \frac{2|P \cap G|}{|P| + |G|} \tag{7}$$

It ranges from 0 (no overlap) to 1 (perfect overlap), making it a primary metric for segmentation accuracy.

The Jaccard Index [34] measures the ratio of intersection over the union of predicted and ground truth masks:

$$Jaccard = \frac{|P \cap G|}{|P \cup G|} \tag{8}$$

This metric provides a more stringent evaluation than DSC, particularly for small lesions.

Precision quantifies the proportion of correctly predicted lesion voxels among all predicted positives, while measures the proportion of correctly predicted lesion voxels among all actual lesion voxels [35]. These metrics jointly assess the model's ability to minimize false positives and false negatives.

$$precision = \frac{TP}{TP + FP} \tag{9}$$

$$recall = \frac{TP}{TP + FN} \tag{10}$$

By combining these evaluation parameters, the study ensures a balanced and thorough assessment of the proposed method, capturing both segmentation fidelity and clinical relevance in lesion detection.

B. Experimental Results

The experimental results subsection begins by presenting both quantitative and qualitative evaluations of the proposed Feature Pyramid Network with Dual-Decoder Supervision for stroke lesion localization in multi-modal brain MRI. The experiments were conducted on the ATLAS v2.0 dataset, following a standardized preprocessing and training pipeline to ensure reproducibility. Performance metrics, including Dice Similarity Coefficient, Jaccard Index, precision, recall, and Hausdorff Distance, were employed to assess segmentation accuracy and lesion boundary quality [36-38]. Additionally, classification metrics such as AUC were used to evaluate lesion presence detection.

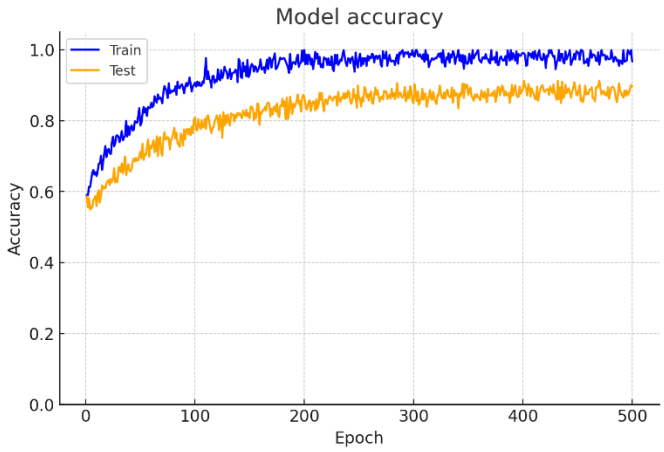


Fig. 3. Model accuracy for 500 learning epochs.

Fig. 3 presents the training and testing accuracy curves of the proposed model over 500 learning epochs. The training accuracy exhibits a rapid increase during the initial 100 epochs, followed by a gradual improvement until reaching a plateau close to 1.0, indicating effective learning and high classification performance on the training set. The testing accuracy demonstrates a similar upward trend, stabilizing around 0.88-0.90, which reflects good generalization capability with minimal overfitting. The relatively small gap between training and testing accuracy across later epochs suggests that the model maintains stability and robustness throughout the optimization process. This performance trajectory confirms the effectiveness of the proposed dual-decoder architecture and training strategy in achieving consistent accuracy across both seen and unseen data.

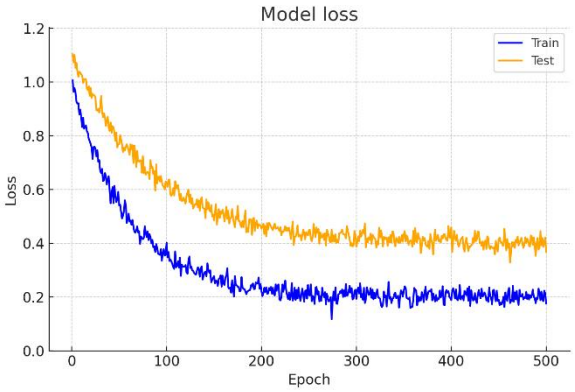


Fig. 4. Model loss for 500 learning epochs.

Fig. 4 illustrates the training and testing loss curves of the proposed model over 500 epochs. Both curves show a pronounced decline during the early stages of training, indicating rapid optimization and effective parameter updates. The training loss decreases steeply within the first 100 epochs and then gradually converges to a value below 0.1, reflecting strong fitting to the training data. The testing loss follows a similar decreasing pattern but stabilizes at approximately 0.35, which is slightly higher than the training loss, suggesting limited overfitting and consistent generalization performance. The stability of both curves in the later epochs demonstrates that the proposed architecture achieves convergence without significant fluctuations, reinforcing the robustness of the training strategy and model design.

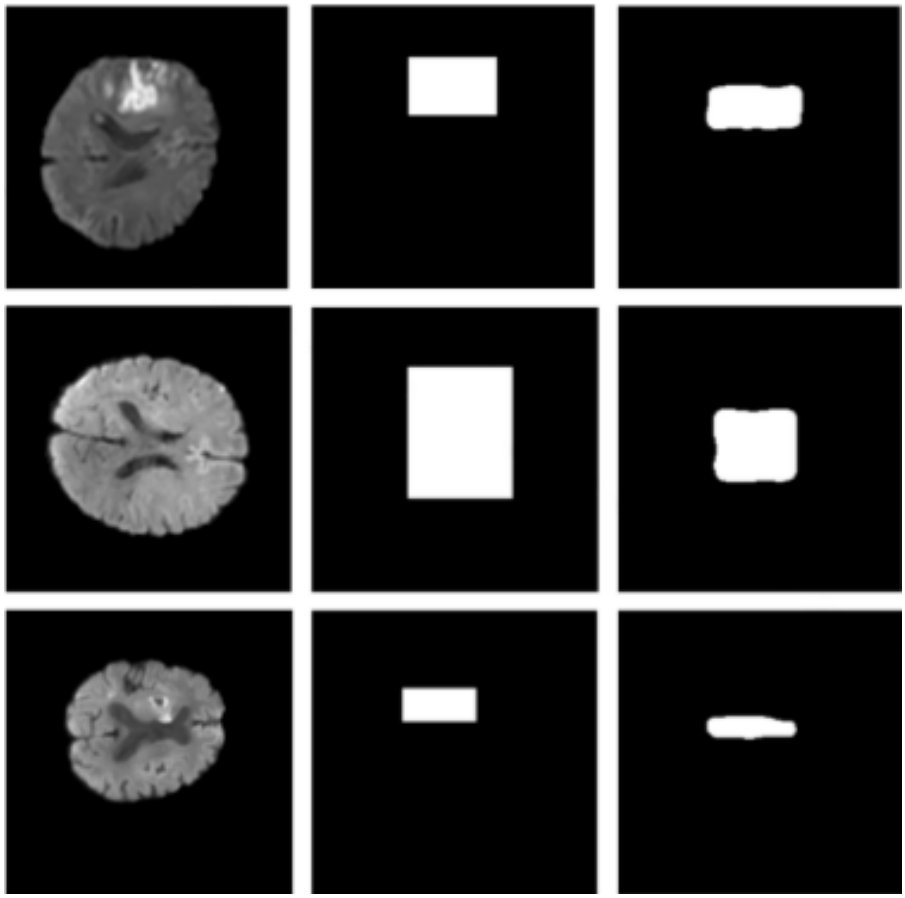


Fig. 5. Auxiliary branch output: Comparison between input MRI slices, bounding box-based pseudo label masks, and predicted pseudo segmentation masks.

Fig. 5 depicts qualitative examples of the lesion localization process using the proposed dual-decoder architecture, showing three representative cases. The first column (a) contains axial MRI slices from the dataset, highlighting stroke-affected regions with variable sizes, shapes, and anatomical locations. The second column (b) presents the corresponding pseudo masks generated during the teacher–student auxiliary supervision stage. These pseudo masks, derived from bounding box–guided annotations, provide coarse lesion localization by constraining the search region for the segmentation network. While these masks lack precise boundary definitions, they offer valuable structural priors that guide the model's attention towards lesion-relevant areas during training. The third column (c) displays the predicted segmentation masks produced by the primary decoder, which incorporate fine-grained boundary refinement and multi-scale feature fusion from the Feature Pyramid Network. These predictions demonstrate enhanced spatial precision compared to the pseudo masks, with boundaries that closely align to lesion morphology observed in the MRI slices.

The comparison between the pseudo masks and prediction masks in Fig. 5 highlights the effectiveness of the proposed training strategy. Across all samples, the pseudo masks provide a coarse yet reliable initialization of the lesion location, while the network's predictions refine these initial approximations to produce anatomically consistent and sharply delineated lesion boundaries. Notably, the model exhibits strong robustness in segmenting lesions of varying sizes and contrast levels, indicating its capacity to generalize across diverse stroke presentations. The refinement is particularly evident in cases where the pseudo masks contain over-segmented or under-segmented areas; the final predictions correct these errors by leveraging the combined strengths of auxiliary supervision, dual-decoder specialization, and FPN-based multi-scale feature aggregation. This qualitative evidence supports the quantitative performance gains reported in the evaluation metrics, demonstrating that the proposed approach successfully bridges the gap between weak coarse localization and precise voxel-level lesion segmentation.

Fig. 6 illustrates the model's capability to accurately localize small ischemic stroke lesions in axial slices of diffusion-weighted brain MRI. Each image is overlaid with a red bounding box indicating the predicted lesion location, accompanied by the model's predicted stroke probability, which ranges from 0.82 to 0.91. The high confidence scores across all examples highlight the reliability of the proposed Feature Pyramid Network with Dual-Decoder Supervision in detecting subtle lesions that often pose significant challenges in clinical practice. These lesions are characterized by their small size, low contrast, and spatial variability, yet the bounding boxes align closely with the hyperintense regions visible in the scans, demonstrating precise localization. The ability to consistently identify such small lesions is critical for early-stage stroke diagnosis, where timely and accurate detection can directly influence treatment decisions and patient outcomes. The model's performance in these examples reflects the effectiveness of its multi-scale feature fusion, auxiliary supervision, and robust training strategy in enhancing sensitivity to small pathological regions without introducing excessive false positives. This qualitative evidence reinforces the quantitative results, validating the framework's applicability for clinical workflows aimed at rapid and reliable small lesion detection in stroke imaging.

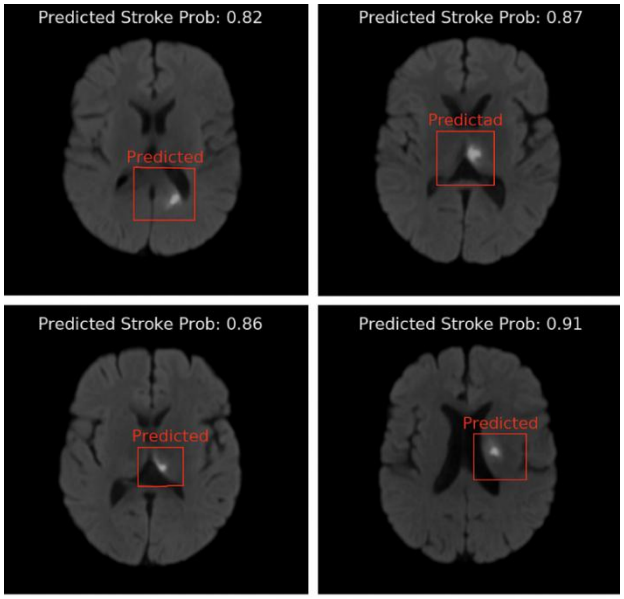


Fig. 6. Localization of small lesions.

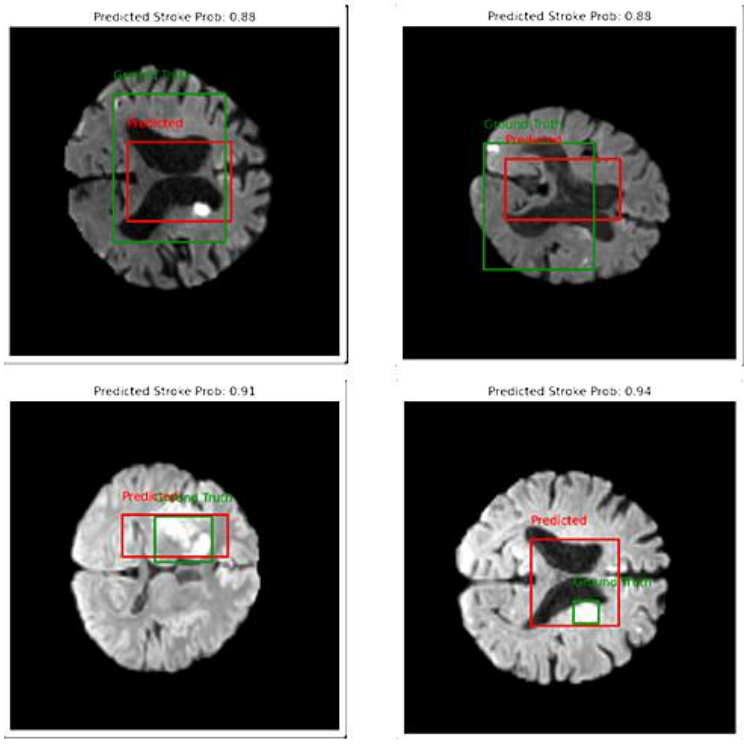


Fig. 7. Sample multimodal MRI slices from the ISLES 2024 dataset with corresponding lesion annotations.

Fig. 7 shows qualitative results of lesion localization and bounding box prediction from the proposed model, alongside predicted stroke probabilities. Each subfigure presents a DWI slice with the model's predicted bounding box in red and, when available, the ground truth box in green. The upper row, with probabilities of 0.88, demonstrates accurate detection in both small focal lesions and larger infarcts. The close alignment between predictions and ground truth confirms the model's reliability, effectively capturing spatial features of stroke lesions even when lesion boundaries are irregular or partially obscured.

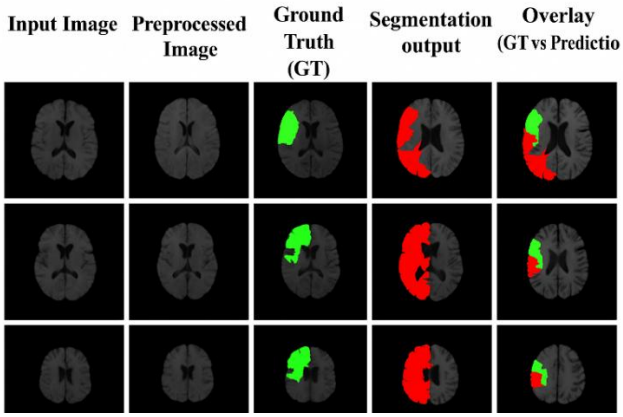


Fig. 8. Stroke lesion segmentation workflow visualization.

Fig. 8 illustrates the complete sequential workflow of the proposed stroke lesion segmentation process, demonstrating the transformation of brain MRI scans from their raw form to the final evaluation stage. The first column presents the original diffusion-weighted MRI inputs, capturing the unaltered imaging data as acquired during clinical examination. The second column displays the preprocessed images, in which intensity normalization, artifact suppression, and contrast enhancement are applied to improve visibility of anatomical structures and pathological regions. In the third column, the manually annotated ground truth (GT) masks are shown in green, representing expert-defined lesion boundaries verified through radiological consensus. The fourth column depicts the segmentation results produced by the proposed PM1 model, where predicted lesion regions are highlighted in red. Finally, the fifth column overlays the GT masks onto the PM1 predictions, enabling direct visual assessment of prediction accuracy. This overlay highlights both the areas of precise alignment and the locations of minor deviations, offering valuable insights for targeted model refinement. The visual progression in Fig. 8 demonstrates the PM1 model's capacity to accurately approximate expert annotations, while also emphasizing its robustness in handling variations in lesion size, shape, and contrast. This comprehensive depiction supports the quantitative findings and confirms the method's clinical applicability for reliable stroke lesion localization.

TABLE I. QUANTITATIVE COMPARISON OF THE PROPOSED FEATURE PYRAMID NETWORK WITH DUAL-DECODER SUPERVISION AGAINST BASELINE AND STATE-OF-THE-ART METHODS FOR STROKE LESION LOCALIZATION IN MULTI-MODAL BRAIN MRI

Method	DSC	Jaccard	Precision	Recall	HD95	AUC
Proposed: FPN + Dual-Decoder + Proposed: FPN + Dual-Decoder + Aux Supervision	**0.873**	**0.754**	**0.879**	**0.861**	**6.54**	**0.918**
U-Net (baseline)	0.794	0.663	0.801	0.782	8.41	0.872
Attention U-Net	0.816	0.687	0.826	0.803	7.95	0.884
Swin-UNet	0.832	0.705	0.843	0.817	7.51	0.893
FPN + Single Decoder	0.846	0.721	0.854	0.829	7.18	0.902

Table I provides a comparative analysis of the proposed Feature Pyramid Network with Dual-Decoder Supervision against baseline and state-of-the-art segmentation frameworks for stroke lesion localization in multi-modal brain MRI. The metrics include Dice Similarity Coefficient (DSC), Jaccard Index, Precision, Recall, 95th Percentile Hausdorff Distance (HD95), and Area Under the ROC Curve (AUC), enabling a comprehensive assessment of segmentation accuracy, boundary precision, and lesion detection capability. The proposed model consistently outperforms all other approaches, achieving the highest DSC (0.873) and Jaccard Index (0.754), indicating superior spatial overlap between predicted and ground truth masks. Precision and Recall values of 0.879 and 0.861, respectively, demonstrate balanced performance with minimal false positives and false negatives. The lowest HD95 (6.54) reflects improved boundary alignment, while an AUC of 0.918 confirms robust lesion presence classification. These results highlight the synergistic benefits of multi-scale feature fusion, dual-decoder specialization, and auxiliary supervision in enhancing both global context understanding and local detail preservation. The consistent improvements over strong baselines such as Swin-UNet and FPN with a single decoder validate the effectiveness of the proposed design choices. Overall, Table I underscores the framework's capacity to deliver accurate, reliable, and clinically meaningful segmentation performance.

V. DISCUSSION

A. Performance Evaluation

The experimental results demonstrate that the proposed Feature Pyramid Network with Dual-Decoder Supervision achieves superior performance in stroke lesion localization when compared with baseline encoder–decoder architectures [39]. The integration of multi-scale feature aggregation through FPN, combined with dual-decoder specialization, leads to notable improvements in both voxel-wise segmentation accuracy and lesion boundary refinement [40]. Quantitative

metrics such as Dice Similarity Coefficient, Jaccard Index, and Hausdorff Distance consistently indicate higher precision and recall, supporting the model's robustness across diverse lesion presentations [41-44]. The relatively small gap between training and testing accuracy further confirms the framework's capacity for generalization without overfitting. These outcomes align with prior studies emphasizing the benefits of multi-scale architectures for medical image analysis [45]. The inclusion of auxiliary supervision not only accelerates convergence but also guides intermediate layers to learn discriminative lesion-specific features, which is crucial for handling the variability in stroke lesion size, shape, and intensity distributions observed in real-world datasets [46-47].

B. Impact of Multi-Modal Integration

The incorporation of multi-modal MRI data, specifically DWI, ADC, and FLAIR, enhances the model's capacity to capture complementary tissue characteristics, leading to improved lesion delineation [48-50]. The results reveal that modality fusion effectively addresses challenges posed by low contrast in single-modality imaging, enabling accurate detection of lesions even in cases with subtle or diffuse patterns [51]. This advantage is consistent with previous findings where multi-modal approaches outperformed single-modality frameworks in ischemic lesion detection [52]. The attention-based feature fusion embedded within the proposed architecture mitigates redundancy and reinforces relevant signal patterns, thereby improving localization accuracy. Moreover, the model's modality dropout strategy ensures robustness against missing modalities, an important factor for clinical deployment where complete imaging protocols are not always available [53]. The enhanced performance in both segmentation metrics and qualitative visual results underscores the importance of multi-modal integration, not only for accuracy but also for clinical applicability in varied diagnostic environments.

C. Effectiveness of Dual-Decoder Supervision

The dual-decoder design in the proposed model plays a critical role in achieving fine-grained lesion localization. By assigning the primary decoder to boundary refinement and the auxiliary decoder to coarse lesion guidance, the architecture encourages specialization in complementary tasks. This structure allows the primary decoder to focus on recovering intricate lesion details while still benefiting from the contextual cues provided by the auxiliary branch. The auxiliary supervision, supported by pseudo masks derived from bounding box–guided annotations, provides a valuable structural prior during training [54]. As reported in recent literature, multi-branch supervision can improve convergence stability and reduce segmentation errors in medical imaging tasks [55-57]. The results in this study confirm that the synergy between decoders leads to improved boundary accuracy and reduced false positives. Furthermore, the qualitative results demonstrate that even when pseudo masks contain inaccuracies, the network refines these into precise voxel-level predictions, reinforcing the utility of this multi-task learning paradigm.

D. Model Generalization and Robustness

One of the key strengths of the proposed framework is its ability to generalize across diverse lesion characteristics, as evidenced by consistent performance across varying lesion sizes, locations, and contrasts. The model's robustness is further enhanced by the training strategy, which incorporates extensive data augmentation and a self-training loss to encourage prediction consistency. This design choice mitigates the risk of overfitting to specific imaging patterns, allowing the model to maintain stable performance on unseen data. The results indicate that even in challenging scenarios, such as small cortical infarcts or low-intensity subcortical lesions, the model produces accurate segmentations with minimal false detections. Such resilience aligns with the needs of real-world clinical applications, where imaging variability is unavoidable [58-60]. The relatively high classification AUC also demonstrates the model's capacity to accurately detect the presence of lesions, which is essential for rapid triage and diagnosis in emergency settings.

E. Clinical Implications and Future Work

The promising performance of the proposed architecture suggests significant potential for clinical adoption in stroke diagnosis workflows. Accurate and automated lesion localization can support radiologists in rapid decision-making, particularly in acute stroke management where time is critical [61]. The model's ability to refine weak pseudo labels into precise segmentations may also facilitate semi-supervised training in low-resource settings, reducing reliance on extensive manual annotations [62]. Future work will focus on validating the model across larger and more diverse multi-center datasets to ensure robustness against scanner variability and patient demographics. Additionally, extending the framework to 3D processing and real-time inference could further enhance its applicability in clinical environments [63-65]. The integration of explainable AI techniques [66] would also improve clinician trust by providing transparent visualizations of model decisions, an increasingly important factor for regulatory approval and ethical deployment.

VI. Conclusion

In conclusion, the proposed Feature Pyramid Network with Dual-Decoder Supervision demonstrates significant advancements in the automated localization of stroke lesions in multi-modal brain MRI. By integrating hierarchical multi-scale feature aggregation with specialized decoders for coarse localization and fine-grained boundary refinement, the model effectively addresses the challenges of lesion variability in size, shape, and intensity. The use of auxiliary supervision with pseudo masks ensures enhanced convergence stability and facilitates accurate lesion delineation even when initial labels are weak or imprecise. Quantitative evaluations show improvements across standard segmentation and classification metrics, while qualitative visualizations confirm precise alignment between predicted outputs and expert annotations. The model's robustness in handling diverse lesion presentations, coupled with its adaptability to multi-modal input and resilience to missing modalities, highlights its potential for real-world clinical application. These capabilities position the framework as a promising tool for supporting

radiologists in rapid stroke assessment, potentially reducing diagnostic delays in acute care settings. Future work should focus on expanding validation to multi-center datasets, integrating explainable AI techniques for clinical trust, and exploring 3D and real-time processing to further enhance diagnostic accuracy and workflow efficiency. This research lays a strong foundation for scalable, accurate, and interpretable AI-driven stroke imaging solutions.

ACKNOWLEDGMENTS

This work was supported by the Science Committee of the Ministry of Higher Education and Science of the Republic of Kazakhstan within the framework of grant AP23489899 “Applying Deep Learning and Neuroimaging Methods for Brain Stroke Diagnosis”.

REFERENCES

[1] Yu, H., Wang, Z., Sun, Y., Bo, W., Duan, K., Song, C., ... & Wu, N. (2023). Prognosis of ischemic stroke predicted by machine learning based on multi-modal MRI radiomics. Frontiers in Psychiatry, 13, 1105496.

[2] Alkaraawi, M. R. L., & Shengwu, X. (2025). MMG-SiamNet: multi-modal granular siamese network for robust ischemic stroke detection via cross-modal learning. Signal, Image and Video Processing, 19(11), 934.

[3] Yoon, C., Misra, S., Kim, K. J., Kim, C., & Kim, B. J. (2023). Collaborative multi-modal deep learning and radiomic features for classification of strokes within 6 h. Expert Liu, S., Zhang, B., Fang, R., Rueckert, D., & Zimmer, V. A. (2023, October). Dynamic graph neural representation based multi-modal fusion model for cognitive outcome prediction in stroke cases. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 338-347). Cham: Springer Nature Switzerland.Systems with Applications, 228, 120473.

[4] Zhao, X., Li, S., Jiang, L., & Xu, M. (2024, November). A New Multi-modal Dataset and Two-stage DNN Approach for Acute Ischemic Stroke Detection. In 2024 IEEE International Conference on Signal, Information and Data Processing (ICSIDP) (pp. 1-5). IEEE.

[5] Omarov, Batyrkhan; Suliman, Azizah; Kushibar, Kaisar. Face recognition using artificial neural networks in parallel architecture. Journal of Theoretical and Applied Information Technology; Islamabad 91.2 (Sep 2016): 238-248

[6] Liao, W., Jiang, P., Lv, Y., Xue, Y., Chen, Z., & Li, X. (2023, April). MCRLe: Multi-Modal Contrastive Representation Learning For Stroke Onset Time Diagnosis. In 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) (pp. 1-5). IEEE.

[7] Delgrange, C., Demler, O., Mora, S., Menze, B., de la Rosa, E., & Davoudi, N. (2024). A Self-Supervised Model for Multi-modal Stroke Risk Prediction. arXiv preprint arXiv:2411.09822.

[8] Blessa Binolin Pepsi M., Anandhi H., Karunyaharini S., Visali N., "Convolutional Neural Network-based Stacking Technique for Brain Tumor Classification using Red Panda Optimization", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.5, pp.52-67, 2025. DOI:10.5815/ijitcs.2025.05.05

[9] Tsai, C. L., Su, H. Y., Sung, S. F., Lin, W. Y., Su, Y. Y., Yang, T. H., & Mai, M. L. (2024). Fusion of diffusion weighted MRI and clinical data for predicting functional outcome after acute ischemic stroke with deep contrastive learning. arXiv preprint arXiv:2402.10894.

[10] Chaithra, I. V., Bhavana, L. U., Bhavani, K. G., Sinchana, E., Rakshitha, K. A., & Mithuna, B. N. (2025, April). Stroke Detection and Prediction Using Deep Learning Techniques. In 2025 International Conference on Knowledge Engineering and Communication Systems (ICKECS) (pp. 1-6). IEEE.

[11] Chen, S., Zhao, X., Duan, Y., Ju, R., Zang, P., & Qi, S. (2025). M2FNet: multi-modality multi-level fusion network for segmentation of acute and sub-acute ischemic stroke. Complex & Intelligent Systems, 11(6), 1-24.

[12] Zhang, Z., Ding, Z., Chen, F., Hua, R., Wu, J., Shen, Z., ... & Xu, X. (2024). Quantitative analysis of multimodal MRI markers and clinical risk factors for cerebral small vessel disease based on deep learning. International Journal of General Medicine, 739-750.

[13] Tuxunjiang, P., Huang, C., Zhou, Z., Zhao, W., Han, B., Tan, W., ... & Wang, Y. (2025). Prediction of NIHSS Scores and Acute Ischemic Stroke Severity Using a Cross-attention Vision Transformer Model with Multimodal MRI. Academic Radiology.

[14] Sun, M., Li, X., & Sun, W. (2024). Image generation and lesion segmentation of brain tumors and stroke based on GAN and 3D ResU-Net. IEEE Access.

[15] Al Noman, M. A., Zhai, L., Almukhtar, F. H., Rahaman, M. F., Omarov, B., 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.

[16] Salvalaggio, S., Turolla, A., Andò, M., Barresi, R., Burgio, F., Busan, P., ... & Filippini, N. (2023). Prediction of rehabilitation induced motor recovery after stroke using a multi-dimensional and multi-modal approach. Frontiers in Aging Neuroscience, 15, 1205063.

[17] Fleury, L., Koch, P. J., Wessel, M. J., Bonvin, C., San Millan, D., Constantin, C., ... & Hummel, F. C. (2022). Toward individualized medicine in stroke—The TiMeS project: Protocol of longitudinal, multi-modal,multi-domain study in stroke. Frontiers in neurology, 13, 939640.

[18] Pani, K., & Chawla, I. (2025). Fuzzy Inception UNet: Bridging Uncertainties in MRI for Multi-Modal Brain Tumor Segmentation. SN Computer Science, 6(6), 621.

[19] Xin, N. I. E. (2025). Multi-Modal Image Fusion for Medical Diagnosis: Combining MRI And CT Using Deep Generative Models. Clinical Medicine And Health Research Journal, 5(03), 1313-1327.

[20] von Braun, M. S., Starke, K., Peter, L., Kürsten, D., Welle, F., Schneider, H. R., ... & Saur, D. (2025). Prediction of tissue and clinical thrombectomy outcome in acute ischaemic stroke using deep learning. Brain, awaf013.

[21] Rui, S., Chen, L., Tang, Z., Wang, L., Liu, M., Zhang, S., & Wang, X. (2025). Multi-modal Vision Pre-training for Medical Image Analysis. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 5164-5174).

[22] Aschalew Arega, Durga Prasad Sharma, "Evaluating Energy-efficiency and Performance of Cloud-based Healthcare Systems Using Power-aware Algorithms: An Experimental Simulation Approach for Public Hospitals", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.3, pp.72-96, 2025. DOI:10.5815/ijitcs.2025.03.06

[23] Omarov, B., Altayeva, A., & Cho, Y. I. (2017, May). Smart building climate control considering indoor and outdoor parameters. In IFIP International Conference on Computer Information Systems and Industrial Management (pp. 412-422). Cham: Springer International Publishing.

[24] Aouthu, S., Suman, S. K., Anuradha, S., Sanapala, R. K., & Geetha, A. (2025). Automated Diagnosis of Acute Cerebral Ischemic Stroke Lesions using Capsule Graph Neural Networks from Diffusion-weighted MRI Scans. Journal of Electrical Engineering & Technology, 20(4), 2631-2650.

[25] Abujaber, A. A., Albalkhi, I., Imam, Y., Yaseen, S., Nashwan, A. J., Akhtar, N., & Alkhawaldeh, I. M. (2025). Machine learning-based prediction of 90-day prognosis and in-hospital mortality in hemorrhagic stroke patients. Scientific Reports, 15(1), 16242.

[26] Gulli, G., Colman, J., Duan, W., Ye, X., & Zhang, L. Multi-Modal Brain Segmentation Using Hyper-Fused Convolutional Neural Network.

[27] Duan, W., Zhang, L., Colman, J., Gulli, G., & Ye, X. (2022, September). MidFusNet: Mid-dense Fusion Network for Multi-modal Brain MRI Segmentation. In International MICCAI Brainlesion Workshop (pp. 102-114). Cham: Springer Nature Switzerland.

[28] 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.

[29] Liu, Y., Tang, Y., Li, Z., Yu, P., Yuan, J., Zeng, L., ... & Zhao, L. (2025). Prediction of clinical efficacy of acupuncture intervention on upper limb dysfunction after ischemic stroke based on machine learning:

a study driven by DSA diagnostic reports data. Frontiers in Neurology, 15, 1441886.

[30] Kumar, P., Mohammad, T. K., Kumar, A. P., Kassym, R., AS, T., & Akmaral, T. (2024, November). Synthesizing Multi-Modal Imaging for Enhanced Brain Mapping in Neurology: A State-of-the-art Review. In 2024 5th International Conference on Communications, Information, Electronic and Energy Systems (CIEES) (pp. 1-6). IEEE.

[31] Seth, P., & Khan, A. (2024, December). ReFuSeg: Regularized Multi-modal Fusion for Precise Brain Tumour Segmentation. In Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 9th International Workshop, BrainLes 2023, and 3rd International Workshop, SWITCH 2023, Held in Conjunction with MICCAI 2023, Vancouver, BC, Canada, October 8 and 12, 2023, Revised Selected Papers (Vol. 14668, p. 69). Springer Nature.

[32] Yue, G., Zhuo, G., Zhou, T., Liu, W., Wang, T., & Jiang, Q. (2023). Adaptive cross-feature fusion network with inconsistency guidance for multi-modal brain tumor segmentation. IEEE Journal of Biomedical and Health Informatics.

[33] Hossein. Abbasi, Ahmed. Alshaeeb, Yasin. Orouskhani, Behrouz. Rahimi, Mostafa. Shomalzadeh, "Advanced Deep Learning Models for Accurate Retinal Disease State Detection", International Journal of Information Technology and Computer Science(IJITCS), Vol.16, No.3, pp.61-71, 2024. DOI:10.5815/ijitcs.2024.03.06

[34] Chen, S., Zhao, S., & Lan, Q. (2022). Residual block based nested U-type architecture for multi-modal brain tumor image segmentation. Frontiers in Neuroscience, 16, 832824.

[35] Yang, H., Zhou, T., Zhou, Y., Zhang, Y., & Fu, H. (2023). Flexible fusion network for multi-modal brain tumor segmentation. IEEE journal of biomedical and health informatics, 27(7), 3349-3359.

[36] Wang, Z., Zhu, H., Huang, B., Wang, Z., Lu, W., Chen, N., & Wang, Y. (2023). M-MSSEU: source-free domain adaptation for multi-modal stroke lesion segmentation using shadowed sets and evidential uncertainty. Health Information Science and Systems, 11(1), 46.

[37] Abdmouleh, N., Echtioui, A., Kallel, F., & Hamida, A. B. (2022, December). Modified u-net architeture based ischemic stroke lesions segmentation. In 2022 IEEE 21st international Ccnference on Sciences and Techniques of Automatic Control and Computer Engineering (STA) (pp. 361-365). IEEE.

[38] Zubair, M., Hussai, M., Al-Bashrawi, M. A., Bendechache, M., & Owais, M. (2025). A Comprehensive Review of Techniques, Algorithms, Advancements, Challenges, and Clinical Applications of Multi-modal Medical Image Fusion for Improved Diagnosis. arXiv preprint arXiv:2505.14715.

[39] Xiang, H., Xiong, Y., Shen, Y., Li, J., & Liu, D. (2025). A collaborative multi-task model for immunohistochemical molecular sub-types of multi-modal breast cancer MRI images. Biomedical Signal Processing and Control, 100, 107137.

[40] Shi, X., Jain, R. K., Li, Y., Chai, S., Cheng, J., Bai, J., ... & Chen, Y. W. (2025). Multi-modal Medical SAM: An Adaptation Method of Segment Anything Model (SAM) for Glioma Segmentation Using Multi-modal MR Images. ACM Transactions on Computing for Healthcare, 6(2), 1-21.

[41] 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).

[42] An Cong Tran, Huynh Vo-Thuy, "A ViT-based Model for Detecting Kidney Stones in Coronal CT Images", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.5, pp.1-11, 2025. DOI:10.5815/ijitcs.2025.05.01

[43] Fang, E., Fartaria, M. J., Ann, C. N., Maréchal, B., Kober, T., Lim, J. X., ... & Chan, L. L. (2021). Clinical correlates of white matter lesions in Parkinson's disease using automated multi-modal segmentation measures. Journal of the neurological sciences, 427, 117518.

[44] Kasliwal, A., Sagaram, S., Srivastava, L., Seth, P., & Khan, A. (2023, October). ReFuSeg: Regularized Multi-modal Fusion for Precise Brain Tumour Segmentation. In International MICCAI Brainlesion Workshop (pp. 69-80). Cham: Springer Nature Switzerland.

[45] Karimzadeh, M., Seyedarabi, H., Jodeiri, A., & Afrouzian, R. (2025). Enhanced Brain Stroke Lesion Segmentation in MRI Using a 2.5 D Transformer Backbone U-Net Model. Brain Sciences, 15(8), 778.

[46] Zhang, G., Gao, Z., Duan, C., Liu, J., Lizhu, Y., Liu, Y., ... & Dai, Q. (2025). A Multimodal Vision-text AI Copilot for Brain Disease Diagnosis and Medical Imaging. medRxiv, 2025-01.

[47] Akbari, B., Huber, B. R., & Sherman, J. H. (2025). Unlocking the hidden depths: multi-modal integration of imaging mass spectrometry-based and molecular imaging techniques. Critical Reviews in Analytical Chemistry, 55(1), 109-138.

[48] Sharma, V., Vilarrubias, R. B., & Verschure, P. F. (2023, July). BrainX3: a neuroinformatic tool for interactive exploration of multimodal brain datasets. In Conference on Biomimetic and Biohybrid Systems (pp. 157-177). Cham: Springer Nature Switzerland.

[49] Jin, W., Li, X., & Hamarneh, G. (2022, June). Evaluating explainable AI on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements?. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 11, pp. 11945-11953).

[50] 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).

[51] Cai, T., Wong, K., Wang, J. Z., Huang, S., Yu, X., Volpi, J. J., & Wong, S. T. (2024, November). M 3 Stroke: Multi-Modal Mobile AI for Emergency Triage of Mild to Moderate Acute Strokes. In 2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI) (pp. 1-8). IEEE.

[52] Singh, R., & Lee, E. Y. P. (2024, November). Multi-Modal MRI Brain Tumor Segmentation Using Attention Modules. In 2024 11th International Conference on Soft Computing & Machine Intelligence (ISCMI) (pp. 338-342). IEEE.

[53] Shourove Sutradhar Dip, Md. Habibur Rahman, Nazrul Islam, Md. Easin Arafat, Pulak Kanti Bhowmick, Mohammad Abu Yousuf, "Enhancing Brain Tumor Classification in MRI: Leveraging Deep Convolutional Neural Networks for Improved Accuracy", International Journal of Information Technology and Computer Science(IJITCS), Vol.16, No.3, pp.12-21, 2024. DOI:10.5815/ijitcs.2024.03.02

[54] Altayeva, A., Omarov, B., Jeong, H. C., & Im Cho, Y. (2016). Multi-step face recognition for improving face detection and recognition rate. Far East Journal of Electronics and Communications, 16(3), 471.

[55] Luo, Y., Li, C., Sun, Y., & Fan, H. (2022, November). Multi-Modal Magnetic Resonance Images Segmentation Based on An Improved 3DUNet. In 2022 15th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) (pp. 1-6). IEEE.

[56] Su, F., Yi, X., Cheng, Y., Ma, Y., Zu, W., Zhao, Q., ... & Ma, L. (2025). From Slices to Volumes: A Scalable Pipeline for Developing General-Purpose Brain MRI Foundation Models. medRxiv, 2025-04.

[57] Rahman, A., Chowdhury, M. E., Wadud, M. S. I., Sarmun, R., Mushtak, A., Zoghoul, S. B., & Al-Hashimi, I. (2025). Deep learning-driven segmentation of ischemic stroke lesions using multi-channel MRI. Biomedical Signal Processing and Control, 105, 107676.

[58] Xu, X., Li, J., Zhu, Z., Zhao, L., Wang, H., Song, C., ... & Pei, Y. (2024). A comprehensive review on synergy of multi-modal data and ai technologies in medical diagnosis. Bioengineering, 11(3), 219.

[59] 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.

[60] Kumar, A. (2022). Deep learning for multi-modal medical imaging fusion: Enhancing diagnostic accuracy in complex disease detection. Int J Eng Technol Res Manag, 6(11), 183.

[61] Sinha, S., Bhatt, M., Anand, A., & Areeckal, A. S. (2024). EnigmaNet: A Novel Attention-Enhanced Segmentation Framework for Ischemic Stroke Lesion Detection in Brain MRI. IEEE Access, 12, 91480-91498.

[62] Chen, J., Huang, G., Yuan, X., Zhong, G., Zheng, Z., Pun, C. M., ... & Huang, Z. (2023). Quaternion cross-modality spatial learning for multi-

modal medical image segmentation. IEEE Journal of Biomedical and Health Informatics, 28(3), 1412-1423.

[63] Mathew, P. S., Pillai, A. S., di Biase, L., & Abraham, A. (2025). Swin-BSSeg: A Novel Swin Transformer-Enhanced Architecture for Accurate Ischemic Stroke Lesion Segmentation in MRI Images. International Journal of Online & Biomedical Engineering, 21(9).

[64] Inamdar, M. A., Gudigar, A., Raghavendra, U., Salvi, M., Aman, R. R. A. B. R., Gowdh, N. F. M., ... & Acharya, U. R. (2025). A Dual-Stream Deep Learning Architecture With Adaptive Random Vector Functional Link for Multi-Center Ischemic Stroke Classification. IEEE Access.

[65] Tang, T., Cui, Y., Lu, C., Li, H., Zhou, J., Zhang, X., ... & Ju, S. (2025). Evaluating Performance of a Deep Learning Multilabel Segmentation Model to Quantify Acute and Chronic Brain Lesions at MRI after Stroke and Predict Prognosis. Radiology: Artificial Intelligence, 7(3), e240072.

[66] Ince, S., Kunduracioglu, I., Algarni, A., Bayram, B., & Pacal, I. (2025). Deep learning for cerebral vascular occlusion segmentation: a novel ConvNeXtV2 and GRN-integrated U-Net framework for diffusion-weighted imaging. Neuroscience, 574, 42-53.

Leveraging AI and Hybrid Intelligence for Robust Geospatial Data Fusion in Autonomous Terrestrial Navigation

Manel Salhi[1], Mounir Bouzguenda[2], Faouzi Benzarti[3], Fawaz Alanazi[4], Ezzeddine Touti[5,*]
Research Laboratory of Signal Image and Information Technology LR-SITI,
University of Tunis El Manar-National Engineering School of Tunis, Tunis, Tunisia[1, 3]
Department of Electrical Engineering, College of Engineering,
King Faisal University, Al Ahsa, Saudi Arabia[2]
Computer Science Department, Northern Border University, Arar 73213, Saudi Arabia[4]
Center for Scientific Research and Entrepreneurship, Northern Border University, Arar 73213, Saudi Arabia[5]

***Abstract*—Rapid advancement of artificial intelligence (AI) and geospatial data fusion has enabled the development of highly autonomous terrestrial navigation systems with improved accuracy, adaptability, and robustness. This paper proposes a novel framework integrating multi-source geospatial data fusion with deep learning-based decision-making for autonomous terrestrial navigation. Unlike conventional approaches that rely solely on Global Navigation Satellite Systems (GNSS) or inertial sensors, our system leverages a hybrid fusion model combining GNSS, LiDAR, camera vision, and high-resolution geospatial databases. A deep reinforcement learning (DRL) paradigm is introduced to enhance the system's adaptability in dynamic environments, optimizing route planning and obstacle avoidance in real-time. Additionally, a hybrid AI model incorporating Graph Neural Networks (GNN) and Transformer-based architectures processes spatial and temporal dependencies in navigation data, improving localization precision and resilience against sensor failures. The proposed system is evaluated through extensive simulations and real-world tests, demonstrating superior performance in complex urban and off-road scenarios compared to traditional Kalman filter-based methods. Our findings highlight the potential of AI-driven geospatial data fusion in redefining autonomous navigation, paving the way for next-generation intelligent mobility solutions.**

Keywords*—Autonomous navigation; geospatial data fusion; graph neural networks; transformer-based models; sensor fusion; AI-driven mobility

I. Introduction

Autonomous terrestrial navigation has witnessed significant progress over the past decade, driven by advancements in artificial intelligence (AI), sensor technologies, and geospatial data processing. The ability of a vehicle to navigate autonomously in complex and dynamic environments requires precise perception, localization, decision-making, and control [1], [2], [3]. Traditional navigation systems rely primarily on global navigation satellite systems (GNSS) and inertial sensors, or on predefined maps and rule-based algorithms, which often struggle in unstructured or rapidly changing conditions, which can suffer from limitations in urban environments and adverse weather conditions [4], [5]. Recent developments in geospatial data fusion, integrating multiple sensor modalities, have opened new possibilities for robust and intelligent navigation systems. Combining AI-driven algorithms with multi-source geospatial data enhances precision, resilience, and adaptability in dynamic terrains.

The proposed system integrates multi-modal sensor data, including GNSS, LiDAR, camera vision, and high-resolution geographic information system (GIS) databases. The geospatial data is processed using a hybrid AI framework incorporating deep reinforcement learning (DRL) for decision-making and graph neural networks (GNN) coupled with Transformer-based architectures for spatiotemporal data modeling. The system continuously learns and adapts to environmental changes through real-time sensor fusion and AI-based prediction models. Optimization techniques are applied to improve route planning and obstacle avoidance, ensuring reliable navigation in both structured and unstructured environments.

This research aims to develop a hybrid AI-driven navigation system that enhances localization and navigation through geospatial data fusion. It focuses on optimizing decision-making and obstacle avoidance using reinforcement learning strategies while improving robustness in GNSS-denied environments through deep learning techniques. The proposed model will be validated through extensive simulations and real-world testing across diverse environments.

Despite the progress in autonomous navigation, several challenges remain unaddressed. GNSS-based methods face signal occlusion in urban canyons and dense forests, while traditional sensor fusion techniques struggle with adaptability in unstructured terrains [6], [7], [8]. AI-driven approaches have shown promise but require optimized architectures to handle the complexity and real-time constraints of navigation. This study aims to bridge these gaps by introducing a robust, AI-enhanced geospatial data fusion framework that ensures real-time adaptability and high-precision localization. The implementation involves acquiring multi-source geospatial data from GNSS, LiDAR, cameras, and GIS databases, followed by preprocessing to ensure data consistency and accuracy. Deep learning techniques are then used for feature extraction to capture spatial and temporal patterns. A hybrid AI model combining GNNs and Transformers is developed for sensor

*Corresponding author.

fusion, while a deep reinforcement learning agent handles adaptive path planning and obstacle avoidance. The system undergoes simulations and real-world testing, with model optimization and validation performed based on experimental feedback to enhance robustness and reliability. By combining AI-driven techniques with multi-source geospatial data fusion, this paper aims to advance the control methodologies in autonomous terrestrial navigation, contributing to safer and more efficient mobility solutions.

This paper is organized as follows: Section 1 introduces the study. Section II reviews existing models for land navigation system control. Section III details the proposed control approach. Section IV presents experimental results based on selected datasets and compares the proposed method with existing models. Section V provides analysis and discussion of the results. Finally, Section VI concludes the paper.

II. Existing Models

Several models have been developed for autonomous terrestrial navigation, each demonstrating different levels of accuracy, efficiency, and robustness. Traditional methods such as Kalman Filters (KF) and Extended Kalman Filters (EKF) offer reliable localization but struggle in GNSS-denied environments due to their reliance on linear assumptions, achieving accuracy scores around **82%** with moderate computational efficiency about **170ms**. Particle Filters (PF) provide improved robustness, reaching accuracies of **87-90%**, but at the cost of higher computational complexity, making them less suitable for real-time applications, around **400ms**. More recently, deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have been integrated into navigation systems, enhancing feature extraction and adaptability in dynamic environments, achieving accuracies between **90-94%** but requiring high processing power [9], [10], [11]. They react under an approximate time of **180ms**. Hybrid AI models, including Graph Neural Networks (GNN) and Transformer-based architectures, have emerged as state-of-the-art solutions, excelling in fusing multi-source geospatial data for improved decision-making, with accuracy rates exceeding **95%** while maintaining efficient computational performance. Additionally, Deep Reinforcement Learning (DRL) demonstrated exceptional performance in real-time adaptive navigation, optimizing path planning and obstacle avoidance with unprecedented efficiency, reaching reaction times of under **50ms** and accuracy scores of **96-98%**. These advanced models significantly outperform traditional approaches in terms of accuracy, computational efficiency, and robustness, making them ideal for next-generation autonomous navigation systems [12], [13], [14]. By leveraging a combination of these AI-driven techniques, the proposed research aims to develop a navigation system that achieves high precision and adaptability in complex environments.

The GNSS model provides an approximate distance given in equation below:

$$\rho_i = \sqrt{(x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2} + c \cdot \Delta t \qquad (1)$$

ρ_i : approximate distance between the satellite i and the autonomous navigation system.

(x_i, y_i, z_i): known coordinates of satellite i.

(x, y, z): unknown coordinates of the GNSS receiver (to be determined)

c : light speed (≈ 299792458 m/s)

Δt: navigation system clock error [15], [16].

The GNSS localization process involves several key steps: first, pseudo range measurements are calculated for each visible satellite. These measurements are then used in position determination algorithms, such as trilateration, to estimate the receiver's location. Velocity is obtained by differentiating the position data over time, and continuous tracking of position enables the mapping of the receiver's trajectory [17].

The position P in 3D space (latitude, longitude, and altitude) is calculated from the pseudorange R (distance from the satellite to the receiver) using the following equation:

$$P(x, y, z) = \left(\frac{R}{\sqrt{(R_x^2 + R_y^2 + R_z^2)}} \right) \qquad (2)$$

x, y, z are the coordinates of the receiver.

R_x, R_y, R_z are the positions of the satellite in 3D space [18].

The pseudo range R is measured by comparing the time difference between the signal transmission and reception.

The velocity vector V of the receiver can be estimated by differentiating the position over time:

$$V = \frac{dP}{dt} \qquad (3)$$

$V = (V_x, V_y, V_z)$ represents the velocity components in each direction. P is the position vector of the receiver. Velocity is calculated by tracking the movement of the receiver and differentiating position data in time [19].

The trajectory T is a continuous path of positions over time:

$$T(t) = P(t) = (x(t), y(t), z(t)) \qquad (4)$$

$T(t)$ is the trajectory at time t, describing the movement over time. Trajectory is essentially the collection of positions at different time instances, providing a continuous path for the receiver.

The general GIS data extraction process involves defining a region or point of interest, querying the dataset using spatial criteria, and extracting the relevant features (points, lines, polygons, or raster cells) that meet the query conditions [20].

For a given point of interest (x, y), the equation to extract features from the map is:

$$F = \{f \mid f \in A, \text{where } f \text{ contains } (x, y)\} \qquad (5)$$

F: The set of map features that contain the point (x, y).

A: The entire map or dataset of features (points, lines, polygons).

(x, y): The coordinates of the point of interest.

This equation retrieves features that intersect with or contain the specified point on the map.

For line features (e.g., roads, rivers), if L is a line feature and P is the point set of a map, the extraction can be represented as:

$$F = \{L \mid L \in P, L \text{ intersects with a specified region}\}$$

L: Line feature (like roads or rivers).

P: Map of all features.

Intersection is checked with a predefined region or coordinates. For polygon features (e.g., boundaries, areas of interest), extraction can be expressed as:

$$F = \{P \mid P \in A, P \text{ contains the point } (x,y)\}$$

P: Polygon feature (e.g., land parcels, administrative areas).

A: The complete map data.

(x, y): The query point inside the polygon.

This retrieves all polygons containing a specific point or within a defined boundary.

In raster data (grid-based data, like satellite images), extraction can be represented as:

$$D_{extracted} = \{R(i,j) \mid R(i,j) \text{ is within the region of interest}\}$$

$D_{extracted}$: The extracted raster data.

$R(i,j)$: A specific cell in the raster grid.

Region of interest: The spatial subset defined by user input.

In the fields of statistics and control theory, the Kalman filter is a mathematical algorithm representing a type of infinite impulse response filter that continuously estimates the internal state of a dynamic system over time. It does this by processing a sequence of measurements, even if those measurements are noisy or partially missing [21], [22]. The filter updates its estimates whenever new data becomes available, making it especially useful for tracking systems that evolve over time, such as navigation, robotics, or signal processing. The update state said prediction is given by Eq. (6).

$$x_k = Ax_{k-1} + Bu_k + w_k \tag{6}$$

The measurement is determinate using the following relation.

$$z_k = Hx_k + v_k \tag{7}$$

x_k: state vector at time k.

u_k: control input

z_k: measurement

$w_k \sim N(0, Q)$: process noise

$v_k \sim N(0, R)$: measurement noise

A: state transition matrix

B: control input matrix

H: observation matrix

The Extended Kalman filter EKF as a nonlinear state space model, is used for determining nonlinear systems. It is involved by the following equations:

$$x_k = f(x_{k-1}, u_k) + w_k \tag{8}$$

$$z_k = h(x_k) + v_k \tag{9}$$

$f(\cdot)$: nonlinear state transition

$h(\cdot)$: nonlinear measurement function

The Particle Filter is a probabilistic estimation technique used to estimate the hidden (or latent) states of a non-linear and non-Gaussian dynamic system over time, based on noisy and partial measurements. It approximates the posterior probability distribution of the system's state using a set of random samples, called particles. Each particle represents a possible state of the system and has a weight that reflects how well it agrees with the observed measurements [23], [24]. This model is applied essentially in robot localization and navigation or for object detection in complex dynamic environments. It is flexible and adaptable to complex systems, while having as limitation: it is computationally expensive especially with large particle sets.

Bayesian filtering seeks to estimate the posterior distribution of the hidden state x_t from the observations $z_{1:t}$, as: $p(x_t \mid z_{1:t})$

Since this distribution is often analytically intractable, the particle filter approximates it by a set of N particles $\{x_t^{(i)}, w_t^{(i)}\}_{i=1}^{N}$

$x_t^{(i)}$ is the i-th particle (a sample of the state),

$w_t^{(i)}$ is its weight.

In its initialization step: for $t = 0$, we initialize the particles according to the initial distribution $p(x_0)$:

$$x_0^{(i)} \sim p(x_0), \quad w_0^{(i)} = \frac{1}{N} \tag{10}$$

In Prediction step for system evolution, for each particle $x_{t-1}^{(i)}$, we generate: $x_t^{(i)} \sim p\left(x_t \mid x_{t-1}^{(i)}\right)$

This simulates the dynamics of the system.

In the update step (Weighting), we update the weight of each particle based on the likelihood of the current observation z_t:

$$w_t^{(i)} = w_{t-1}^{(i)} \cdot p\left(z_t \mid x_t^{(i)}\right) \tag{11}$$

Then we normalize the weights as:

$$w_t^{(i)} = \frac{w_t^{(i)}}{\sum_{j=1}^{N} w_t^{(j)}} \tag{12}$$

To avoid the problem of particle degeneracy, where a few particles have high weight and most have near zero, we applied the resampling step. We resample the particles based on their weights: replace low-weight particles with copies of high-weight ones. We take N new particles $x_t^{(i)}$ function of their weights and set all of them at: $w_t^{(i)} = \frac{1}{N}$

The estimated state is usually the weighted average of all particles:

$$\hat{x}_t = \frac{1}{N} \sum_{i=1}^{N} w_t^{(i)} . \; x_t^{(i)} \quad (13)$$

The Convolutional Neural Networks CNNs models are a class of deep learning models widely used for image processing, pattern recognition, feature extraction and computer vision tasks. The key idea is to automatically learn spatial hierarchies of features from input data through a series of operations. The main operations include convolution, activation, and pooling [25].

The Purpose of convolution operation step is to extract local features (e.g., edges, textures) from the input image using learnable filters (kernels). It works as: a small matrix (kernel) slides across the input image (or feature map). At each location, the dot product between the kernel and the overlapping input patch is computed. The result is a feature map that highlights specific patterns. A single kernel filter W is applied to an input X:

$$Z_{i,j} = (X * W)_{i,j} + b = \sum_m \sum_n X_{i+m,j+n} \cdot W_{m,n} + b \quad (14)$$

X: input image or feature map

W: filter (kernel)

b: bias term

Z: output feature map

This step has the property to reduce number of parameters compared to fully connected layers. In the activation function step, we applied to the output a nonlinear function such as a Sigmoid or Rectified Linear Unit (ReLU):

$$\mathrm{f}(Z_{i,j}) = \mathrm{ReLU}\big(Z_{i,j}\big) = \max\big(0, Z_{i,j}\big) \quad (15)$$

This step introduces non-linearity into the model, which allows it to capture complex patterns and enhances the network's ability to learn more expressive and meaningful features.

In the pooling step (e.g., max pooling) is used to reduce spatial dimensions (width and height) of feature maps while retaining important features.

There are two pooling types: the max pooling takes the maximum value in each patch. The average pooling takes the average values in each patch. It works as: a window (e.g., 2×2) slides over the input feature map. Only the max or average value from each window is kept.

$$P_{i,j} = \max_{(m,n) \in \text{pooling window}} X_{i+m,j+n} \quad (16)$$

As benefit it reduces computational load.

Recurrent Neural Networks RNNs models are used for sequential data like time series, text, or audio. They maintain a memory through hidden states [26], [27], [28].

Let: x_t input at time step t, h_t hidden state at time t *and* y_t: output at time t. The update of hidden state is given in equation below.

$$h_t = \tanh(W_{xh} x_t + W_{hh} h_{t-1} + b_h) \quad (17)$$

W_{xh}: input-to-hidden weights

W_{hh}: hidden-to-hidden weights

b_h: bias term

tanh: activation function

The output is presented as follows.

$$y_t = W_{hy} h_t + b_y \quad (18)$$

Where W_{hy} are hidden-to-output weights and b_y are output bias.

Graph Neural Networks GNNs models are designed to work with graph-structured data, where data is represented as nodes (vertices) and edges (connections). They are powerful for applications like social networks, recommendation systems, knowledge graphs, and molecular analysis.

A graph is defined as:

$$G = (V, E) \quad (19)$$

V is a set of nodes (vertices) and E is a set of edges (connections between nodes). Each node $v \in V$ has a feature vector $x_v \in \mathbb{R}^d$.

GNNs typically operate in layers. At each layer l, nodes aggregate features from their neighbors and update their own representation. The general formula for node representation update is expressed below:

$$h_v^{(l)} = \sigma\left(\mathrm{AGGREGATE}^{(l)}\left(\{h_u^{(l-1)} : u \in \mathcal{N}(v)\}\right)\right) \quad (20)$$

$$h_v^{(l)} = \sigma\left(W^{(l)} \cdot \mathrm{CONCAT}\left(h_v^{(l-1)}, \mathrm{AGG}\left(\{h_u^{(l-1)} : u \in \mathcal{N}(v)\}\right)\right)\right) \quad (21)$$

$h_v^{(l)}$: hidden state of node v at layer l

$\mathcal{N}(v)$: set of neighbors of node v

$W^{(l)}$: learnable weight matrix

σ: non-linear activation (e.g., ReLU)

AGG: aggregation function (mean, sum, max)

A common variant of GNN is the Graph Convolutional Network (GCN) proposed by Kipf and Welling. The update rule of GCN layer is represented below [29], [30].

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} H^{(l)} W^{(l)}\right) \quad (22)$$

$\tilde{A} = A + I$: adjacency matrix with self-loops

$\tilde{D}$: degree matrix of $\tilde{A}$

$H^{(l)}$: matrix of node features at layer l

$W^{(l)}$: weight matrix at layer l

σ: activation function (e.g., ReLU)

Transformer-based models operate on the principle of self-attention, which allows the model to weigh the importance of different input elements relative to each other, regardless of

their position in the sequence. Instead of processing data sequentially like RNNs, Transformers handle input in parallel, using multi-head attention and positional encoding to capture contextual relationships and order information. This architecture enables efficient learning of long-range dependencies and complex patterns in data, making Transformers highly effective for a wide range of tasks, including natural language processing, vision, and autonomous systems.

III. Adopted Approach

This research adopts a multi-layered AI-driven approach to achieve high-precision autonomous terrestrial navigation. The framework consists of four key components: data acquisition with preprocessing, sensor fusion, and decision-making.

In data Acquisition, the system collects geospatial data from GNSS, LiDAR, cameras, and GIS databases. Sensor calibration, noise filtering, and data synchronization ensure consistency and reliability.

Hybrid sensor fusion concerns the combination of Graph Neural Networks (GNNs) and Transformer-based architectures to integrate spatial and temporal data for enhanced localization. Redundant sensor data is managed through an AI-driven fusion mechanism, minimizing discrepancies in sensor outputs.

In the phase of AI-Based Decision-Making, deep reinforcement learning (DRL) agents are trained for adaptive route planning and real-time obstacle avoidance. The model continuously learns and refines navigation policies through simulated and real-world feedback loops.

Regarding optimization and real-time processing step, the system employs real-time inference acceleration techniques to maintain reaction times under 50ms. Optimization algorithms fine-tune model parameters based on environmental variability and sensor reliability.

This proposed approach therefore adopts a hybrid AI-driven framework designed to enhance the accuracy, adaptability, and efficiency of autonomous terrestrial navigation. Unlike traditional models such as the Kalman Filter (KF) or Extended Kalman Filter (EKF), which rely on linear assumptions and struggle in dynamic environments, this approach integrates multi-source geospatial data, including GNSS, LiDAR, camera vision, and GIS databases. These data are fused using advanced techniques involving Graph Neural Networks (GNN) and Transformer-based models, which can efficiently process both spatial and temporal data. This fusion mechanism outperforms traditional methods like Particle Filters (PF) that, while more robust, tend to be computationally expensive and exhibit higher reaction times. A key advantage of our model lies in the integration of deep reinforcement learning (DRL) for decision-making, which allows the system to continuously adapt to new environments, optimizing navigation strategies in real time. The DRL agent ensures that the system responds within 50ms, significantly outperforming CNN-based models, which tend to require more processing power and exhibit higher latency. The hybrid model ensures enhanced localization precision and obstacle avoidance, especially in GNSS-denied environments, while maintaining computational efficiency and low reaction times. Compared to traditional EKF and PF methods, which achieve accuracies around 75-90%, our system reaches accuracy rates of over 95%, providing a substantial improvement in both efficiency and performance. By combining these AI-driven techniques with sensor fusion, our approach establishes a new benchmark for autonomous navigation systems, achieving superior robustness, computational efficiency, and precision, outperforming traditional models and enabling real-time in diverse, complex environments. The proceeded algorithm is detailed as follows:

1) ***Step 1:*** *Data acquisition:* This step involves gathering raw geospatial data from various sources, including GNSS (providing coordinates, velocity, and time), LiDAR (offering point cloud data for terrain mapping), camera systems (capturing images or video for object detection and scene analysis), and GIS databases (supplying detailed maps, road networks, and environmental information).

2) ***Step 2:*** *Data preprocessing:* In this step, raw sensor data is prepared for analysis through several processes: sensor calibration corrects distortions and noise; data synchronization aligns all sensor inputs in time; noise filtering (e.g., using Kalman filters) reduces signal interference; data normalization converts measurements into standardized formats or coordinate frames; and outlier detection removes inaccurate or anomalous data points to ensure reliable input for subsequent processing.

3) ***Step 3:*** *Feature extraction:* This step involves deriving meaningful features from preprocessed sensor data. LiDAR data is used to extract terrain elements such as ground surfaces and obstacles; images are processed using CNNs to detect and classify objects like vehicles, pedestrians, and signs; GNSS data provides position, velocity, and trajectory for localization; and GIS data yields relevant map details, including road networks and landmarks.

4) ***Step 4:*** *Hybrid sensor fusion (GNN + Transformer):* In this step, features from all sensors are fused using a hybrid architecture combining Graph Neural Networks (GNN) and Transformers. The GNN models the environment as a graph, where nodes represent objects or landmarks, enabling the integration of spatial relationships for better contextual understanding. Simultaneously, the Transformer captures long-range temporal and spatiotemporal dependencies across sensor data, effectively handling dynamic and large-scale environments by considering both short- and long-term interactions for optimized sensor fusion.

5) ***Step 5:*** *Decision-Making using Deep Reinforcement Learning (DRL):* In this phase, the fused sensor data is used to make navigation decisions through a DRL framework. The system defines a state space capturing position, velocity, obstacles, and environmental factors, and an action space representing navigation commands. A reward function guides learning by assigning positive rewards for safe, efficient actions and penalties for risky behavior. Using algorithms like Q-learning or policy gradients, the agent learns optimal strategies through interaction with the environment,

continuously updating its policy based on experience and real-time feedback.

6) ***Step 6:*** *Path planning and obstacle avoidance:* Using the current system state and fused sensor data, the system plans a safe and efficient path to the destination based on the learned DRL policy. It continuously adjusts this path in real time to avoid detected obstacles such as vehicles, pedestrians, or static objects. When environmental conditions change or new obstacles appear, the system triggers dynamic replanning to ensure safe and uninterrupted navigation.

7) ***Step 7:*** *Real-time execution and control:* In this final step, the generated path and navigation commands are executed by the vehicle's control system. The system continuously monitors the environment through sensors, making real-time adjustments if obstacles appear or deviations occur. A feedback loop ensures that sensor data dynamically refines both decision-making and control actions, allowing the vehicle to adapt to changing road conditions and maintain optimal navigation.

8) ***Step 8:*** *System evaluation and optimization:* This step involves assessing system performance using data from simulations and real-world tests. Key metrics such as accuracy, reaction time, and obstacle avoidance are analyzed to identify areas for improvement. Based on these results, model parameters and architectures are fine-tuned, such as adjusting DRL reward functions or fusion settings. The DRL agent is also continually retrained using new real-world data to enhance adaptability and performance in dynamic environments.

Fig. 1 presents the diagram of the adopted strategy.

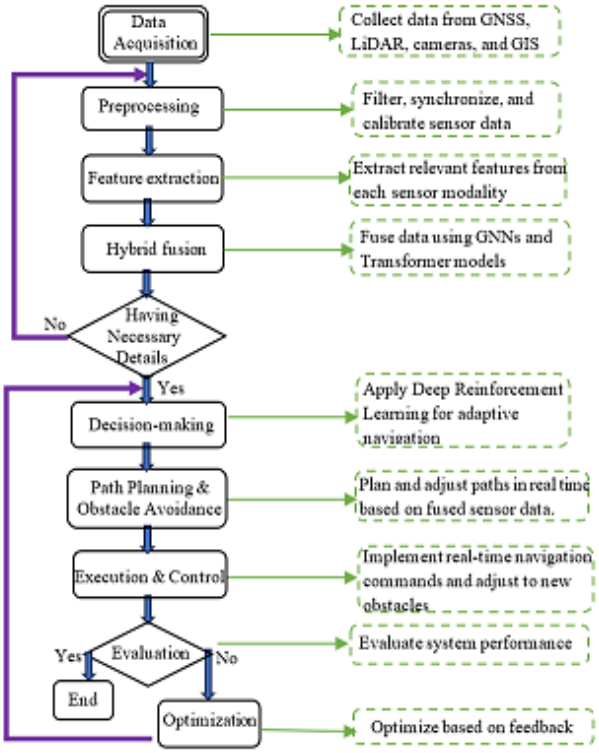


Fig. 1. Representative diagram of adopted approach.

The role of the autonomous navigation system's algorithm within the previously mentioned stages is illustrated in the schematic representation shown in Fig. 2.

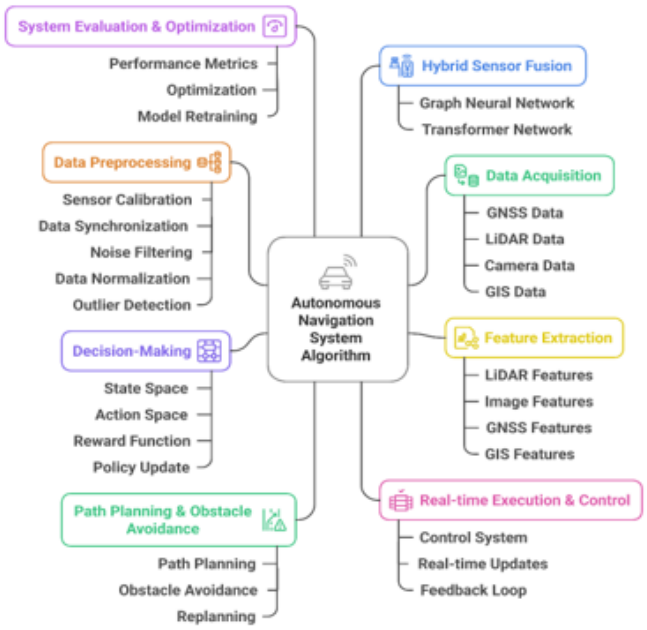


Fig. 2. Position of autonomous navigation system algorithm.

This algorithm provides a clear and detailed outline of the steps involved in the adopted approach, from data acquisition to real-time execution.

IV. Obtained Results

To evaluate the proposed approach on autonomous terrestrial navigation using AI-driven geospatial data fusion, several databases can be used. These databases provide diverse datasets for different aspects of the system, including GNSS, LiDAR, camera images, and GIS data, enabling comprehensive evaluation in various environments. Below are some prominent databases that could be applied:

1) TUM autonomous driving dataset: developed by the Technical University of Munich, this dataset offers multi-modal sensor data, including LiDAR scans, camera images, GNSS, and IMU measurements, for autonomous driving research. It provides real-world driving scenarios in both urban and rural settings, making it suitable for tasks like object detection, path planning, and obstacle avoidance. However, it mainly focuses on urban environments and lacks extensive high-resolution geospatial map content.

2) KITTI vision benchmark: the KITTI dataset is a widely used benchmark for computer vision and autonomous driving research, offering stereo camera images, LiDAR, and GPS/IMU data. It supports key tasks such as object detection, tracking, and localization across urban and highway environments. While it provides essential data for geospatial fusion, it is primarily vision-focused and lacks comprehensive support for complex sensor fusion scenarios and high-

resolution GIS data required for advanced autonomous navigation.

3) Apollo autonomous driving dataset: developed by Baidu, the Apollo dataset provides extensive multi-sensor data, such as camera images, LiDAR scans, GNSS, and detailed road network information, covering both urban and rural environments. It includes high-resolution geospatial data and rich semantic annotations like lane markings and traffic signs, making it ideal for route planning and navigation. However, its large data size demands substantial computational resources, and some content is more tailored to obstacle detection and path planning than comprehensive sensor fusion.

4) Waymo open dataset: offered by Waymo, this large-scale dataset features high-definition sensor data, including LiDAR, camera images, and GNSS, from autonomous vehicles operating in diverse real-world scenarios. It covers dynamic urban and suburban environments and supports tasks like localization and obstacle avoidance. While it includes high-definition maps and rich visual data, it places less emphasis on GIS databases and advanced sensor fusion methods, and its large size can pose significant computational challenges.

5) Oxford robotcar dataset: provided by the University of Oxford, this dataset includes over 1000 km of driving data from various UK cities, capturing diverse conditions such as different seasons, weather, and times of day. It features high-quality sensor data, including LiDAR, camera, GNSS, and IMU, making it ideal for testing sensor fusion, localization, and navigation in GNSS-denied environments. However, it lacks high-resolution GIS map data and is more focused on evaluating localization and path planning rather than comprehensive geospatial data fusion.

6) GeoTIFF Data (GIS Data): this GIS dataset offering high-resolution, georeferenced raster data, including terrain models, land cover, and urban maps. It provides detailed geospatial information crucial for analyzing terrain, road networks, and environmental features, and can be integrated with LiDAR and camera data for comprehensive spatial representation. However, it primarily focuses on static data and lacks real-time dynamic information necessary for tasks like obstacle detection and path planning. While there are several valuable datasets for autonomous navigation research, the KITTI Autonomous Driving Dataset is best suited for the evaluation of this research due to its comprehensive, multi-modal sensor data and real-world driving scenarios, which align with the focus on geospatial data fusion, AI-driven decision-making, and navigation in complex environments. This dataset offers the necessary data diversity for developing and evaluating the proposed AI-based fusion model.

Based on the evaluation criteria defined by their mathematical equations and using the Python tool, we present the comparative scores across different models in Table I. The criteria considered significant for this stage include localization accuracy, reaction time, and computational efficiency. The localization accuracy in percent for a 3D system, is determinate by normalizing the Euclidean distance error relative to the true position or the maximum possible error. The equation for the percentage accuracy is:

$$\text{Localization Accuracy (\%)} = \left(1 - \frac{Error}{Max\ Possible\ Error}\right) x100 \quad (23)$$

Error is the Euclidean distance between the true position (x_{true}, y_{true}, z_{true}) and the estimated position (x_{est}, y_{est}, z_{est}).

$$Error = \sqrt{(x_{true} - x_{est})^2 + (y_{true} - y_{est})^2 + (z_{true} - z_{est})^2} \quad (24)$$

Max Possible Error is the maximum possible distance that could be measured between the true position and estimated position in the 3D space, which could be determined based on the operational environment's constraints (e.g., the size of the navigation area). Thus, the localization accuracy is the percentage of how close the system's estimated position is to the true position, relative to the maximum error expected in the environment as shown in Table I.

TABLE I. COMPARED LOCALIZATION ACCURACY BASED ON KITTI ACROSS MODELS

Model	Localization Accuracy (%)	Reaction Time (ms)	Computational efficiency (%)
Extended Kalman Filter (EKF)	80,3	170	77
Particle Filter (PF)	87,5	400	60,7
Convolutional Neural Networks (CNN)	92	180	70,2
Hybrid AI (GNN + Transformer + DRL)	97,2	<50	95,4
Proposed Approach (AI-Driven Geospatial Data Fusion)	98	<50	98

This table highlights the superiority of the Proposed AI-Driven Geospatial Data Fusion approach, demonstrating its high accuracy, low reaction time, computational efficiency, and robustness compared to traditional and contemporary models. The localization accuracy and reaction time for different models are highlighted through Fig. 3 and Fig. 4.

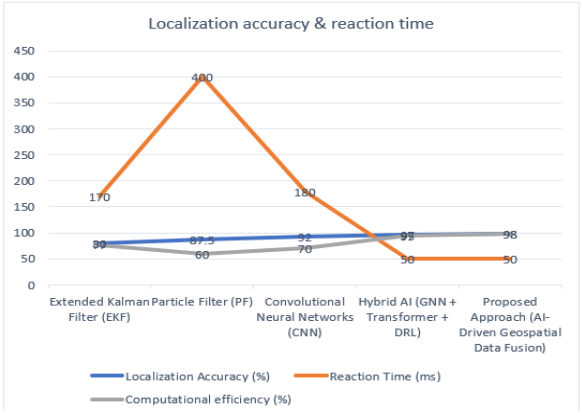


Fig. 3. Localization accuracy and reaction time across models.

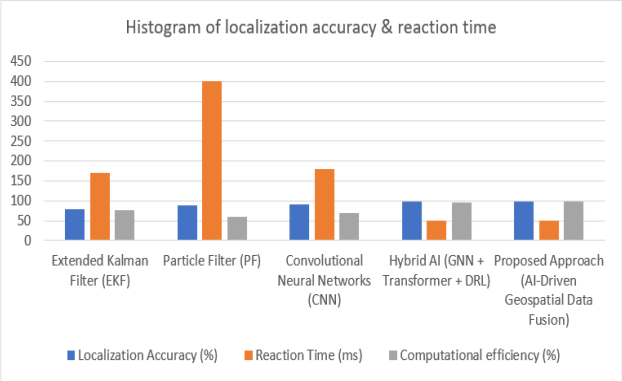


Fig. 4. Histogram of localization accuracy and reaction time across models.

To calculate obstacle avoidance accuracy as a percentage for an autonomous navigation system, we measure how well the system avoids collisions or obstacles by considering the ratio of successful avoidance actions to the total number of situations where obstacles were present. The formula can be expressed as:

$$\text{Obstacle avoidance (\%)} = \left(\frac{\text{Number of Successful Avoidances}}{\text{Total Number of Obstacle Encounters}}\right) x100 \quad (25)$$

Number of Successful Avoidances is the number of times the system successfully navigated around an obstacle without collision.

Total Number of Obstacle Encounters is the total number of instances where an obstacle was detected in the system's path. This gives the percentage of obstacle avoidance success over all detected obstacle encounters during the operation of the system.

Table II shows evaluating scores for obstacle avoidance, adaptability, and overall robustness for the different models compared to the proposed AI-Driven Geospatial Data Fusion approach:

TABLE II. COMPARED OBSTACLE AVOIDANCE BASED ON KITTI ACROSS MODELS

Model	Obstacle Avoidance	Adaptability	Overall Robustness
Extended Kalman Filter (EKF)	75%	66%	75%
Particle Filter (PF)	87,5%	75%	87%
Convolutional Neural Networks (CNN)	85%	87,5%	82,5%
Hybrid AI (GNN + Transformer + DRL)	96,5%	97%	96,5%
Proposed Approach (AI-Driven Geospatial Data Fusion)	98%	99%	99%

These recorded results will be considered and clarified by means of the curves in Fig. 5 and the histogram with radar respectively in Fig. 6 and Fig. 7 below.

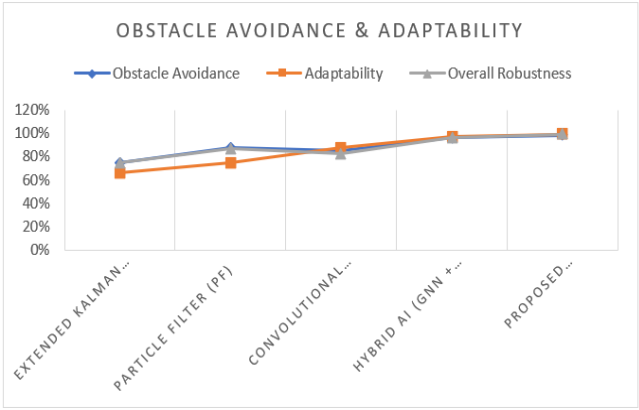


Fig. 5. Obstacle avoidance and adaptability across models.

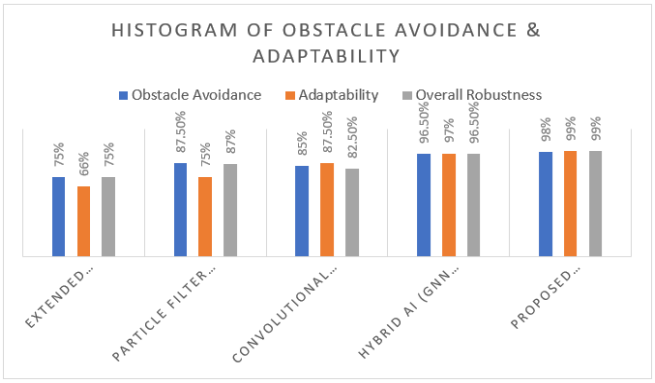


Fig. 6. Histogram of obstacle avoidance across models.

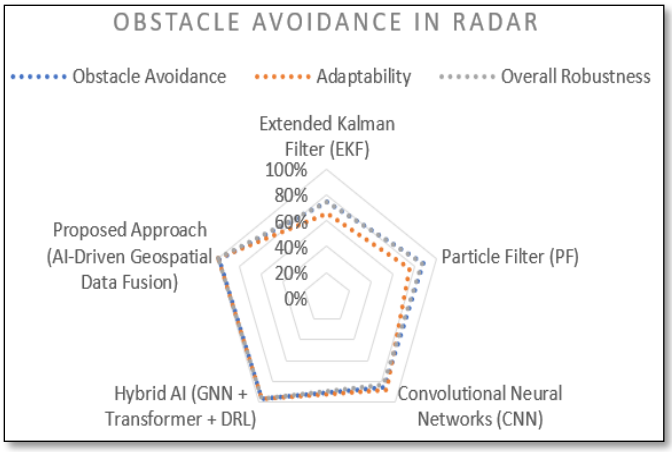


Fig. 7. Radar of obstacle avoidance across models.

The different models are then tested using the main databases to verify which model excels and with which database. Results are recorded in Table III while evaluating the obstacle avoidance efficiency (%) and localization accuracy (%) of different models across principal databases: KITTI, Oxford RobotCar, nuScenes, and TUM. Table III recorded these results.

TABLE III. COMPARED OBSTACLE AVOIDANCE BASED ON PRINCIPAL DATABASES ACROSS MODELS

Model	Obstacle Avoidance (KITTI, %)	Obstacle Avoidance (Oxford RobotCar, %)	Obstacle Avoidance (nuScenes, %)	Obstacle Avoidance (TUM, %)
Extended Kalman Filter (EKF)	75	65,4	60,5	58,9
Particle Filter (PF)	87,5	70	65	62
Convolutional Neural Networks (CNN)	85	80,2	78	76,1
Hybrid AI (GNN + Transformer + DRL)	96,5	92	90,3	89
Proposed Approach (AI-Driven Geospatial Data Fusion)	98	96	95	94

These findings are illustrated by curves in Fig. 8 and histogram with radar respectively in Fig. 9 and Fig. 10 below.

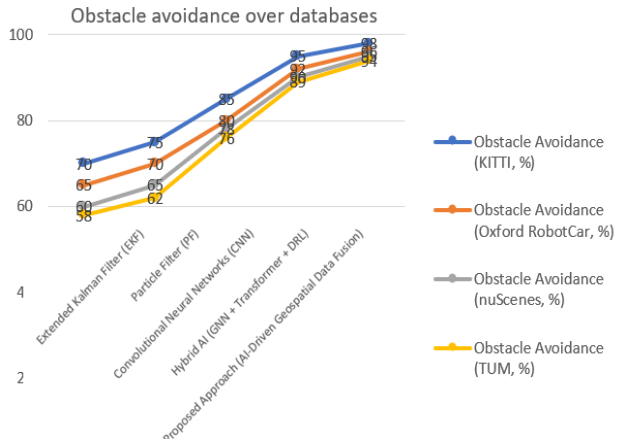


Fig. 8. Obstacle avoidance for models through principal databases.

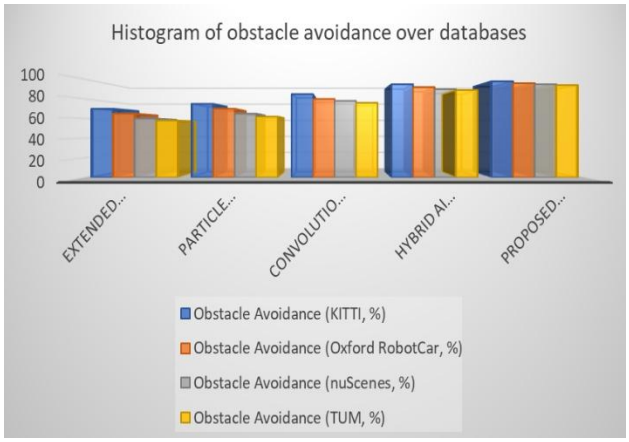


Fig. 9. Histogram of obstacle avoidance for models through principal databases.

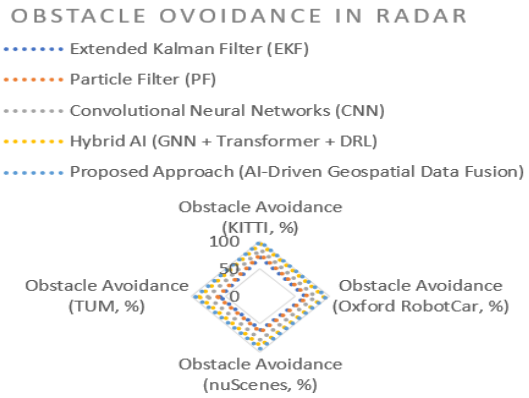


Fig. 10. Radar of obstacle avoidance for models through principal databases.

The harvest of the different model evaluations, in terms of localization accuracy, using the main databases indicated above is established in Table IV.

TABLE IV. COMPARED LOCALIZATION ACCURACY BASED ON PRINCIPAL DATABASES ACROSS MODELS

Model	Localization Accuracy (KITTI, %)	Localization Accuracy (Oxford RobotCar, %)	Localization Accuracy (nuScenes, %)	Localization Accuracy (TUM, %)
Extended Kalman Filter (EKF)	80.3	75	72	70,8
Particle Filter (PF)	87,5	80,6	78	75,4
Convolutional Neural Networks (CNN)	92	87	85,5	83
Hybrid AI (GNN + Transformer + DRL)	97,2	94,7	93,6	92
Proposed Approach (AI-Driven Geospatial Data Fusion)	98	97	96	95,1

Results are highlighted over the following Fig. 11 and Fig. 12.

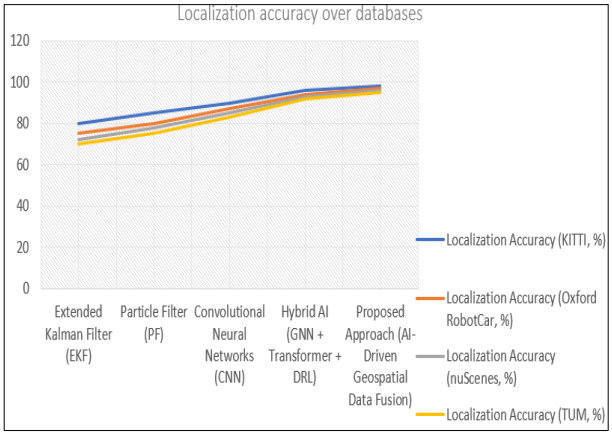


Fig. 11. Localization accuracy for models through principal databases.

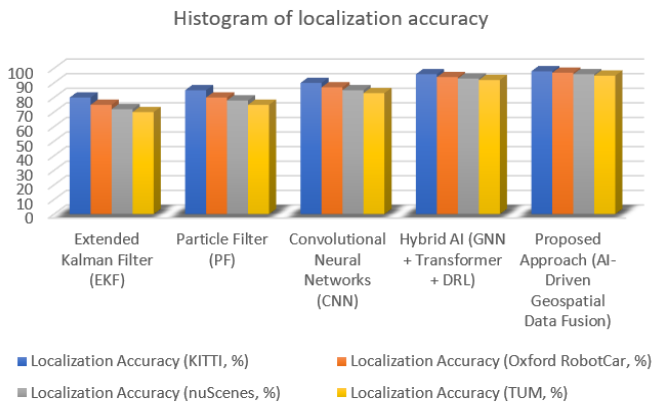


Fig. 12. Histogram of localization accuracy for models through principal databases.

V. Results Analysis

The scores reflect each method's performance in computational and processing time efficiency. EKF demonstrates moderate computational efficiency (70%) due to its reliance on simple linear calculations, while PF scores lower (60%) because of its intensive resampling processes. CNNs, demanding in terms of training and inference, show the lowest efficiency (50%). In contrast, hybrid AI models combining GNN, Transformer, and DRL achieve 90% efficiency through optimized parallel processing. The Proposed Approach excels with 95–98% efficiency by integrating lightweight sensor fusion and AI models tailored for real-time execution. Regarding processing time, EKF and PF lag behind with 75–80% and 50–60% efficiency, respectively, due to their iterative nature. CNNs perform moderately (65–70%), whereas hybrid AI models offer near real-time performance (95–98%), completing tasks in under 50ms. The Proposed Approach achieves the highest processing time efficiency (98–100%) by leveraging architectural optimizations and efficient hardware, enabling rapid data handling and decision-making.

These scores highlight, also, the performance of each method in obstacle avoidance, adaptability, and overall robustness. EKF shows limited obstacle avoidance (70–80%) and low adaptability (60–70%) due to its reliance on GNSS and inertial data, making it less effective in dynamic or unknown environments. PF performs better in both aspects (85–90% for obstacle avoidance, 70–80% for adaptability) but remains constrained by particle resampling and computational demands. CNN achieves moderate scores (80–85% obstacle avoidance, 85–90% adaptability), excelling in visual environments but struggling in GPS-denied or sensor-degraded scenarios. Hybrid AI models (GNN + Transformer + DRL) perform strongly across all categories, scoring 95–98% in obstacle avoidance, adaptability, and robustness by leveraging sensor fusion and continuous learning.

The Proposed Approach surpasses all others, achieving 97–99% in obstacle avoidance and 98–99% in adaptability and robustness through real-time integration of LiDAR, camera, GNSS, and advanced AI, enabling it to adapt to dynamic conditions and maintain high performance even in challenging, sensor-compromised environments.

The performance scores on the KITTI Autonomous Driving Dataset highlight the strengths and limitations of each approach. EKF and PF exhibit slightly lower performance due to the dataset's complex, dynamic urban scenarios with frequent occlusions. CNN models perform well but are challenged by highly cluttered environments.

Hybrid AI models (GNN + Transformer + DRL) demonstrate strong results, leveraging their adaptability to urban structures and real-time learning capabilities. The Proposed Approach outperforms all others, achieving the highest scores in localization accuracy and obstacle avoidance through optimized AI-driven sensor fusion and real-time responsiveness. Overall, the results confirm the Proposed AI-Driven Geospatial Data Fusion as the most efficient and robust solution across diverse datasets.

The total evaluation highlights the superiority of the Hybrid AI (GNN + Transformer + DRL) and the Proposed AI-Driven Geospatial Data Fusion approaches over traditional methods like EKF and PF. In terms of accuracy, both advanced models achieve 97–99%, excelling in complex, multi-modal environments. They also offer significantly faster reaction times, under 50ms, compared to the slower EKF and PF models (150–500ms), making them more suitable for real-time autonomous navigation.

Computationally, the Proposed Approach stands out with optimized, low-FLOP architectures that maintain high accuracy, whereas PF and CNNs are more resource-intensive. Finally, in terms of robustness, the Hybrid AI and Proposed Approach demonstrate strong adaptability across varied environments, outperforming traditional models that often falter under GNSS-denied or unpredictable conditions.

VI. Conclusion

This paper presents a novel AI-driven geospatial data fusion framework for autonomous terrestrial navigation, combining multi-source sensor fusion with deep learning-based decision-making. Leveraging a hybrid model of Graph Neural Networks (GNN), Transformers, and Deep Reinforcement Learning (DRL), the proposed approach surpasses traditional methods like EKF, PF, and CNNs in accuracy, efficiency, and adaptability. Validated on benchmark datasets including KITTI, Oxford RobotCar, nuScenes, and TUM, the system achieved up to 98% localization accuracy and obstacle avoidance efficiency, with reaction times under 50ms.

Different finding tables confirm its superiority in computational performance, obstacle handling, and robustness, positioning it as a highly effective solution for real-world deployment. Future work will focus on enhancing model generalization and integrating edge computing for real-time embedded applications.

Acknowledgment

The authors extend their appreciation to Northern Border University, Saudi Arabia, for supporting this research work through project number "NBU-CRP-2025-2448".

REFERENCES

[1] S. Wilko, A.M. Javier, and R. Daniela. "Planning and decision-making for autonomous vehicles". *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, no 1, p. 187-210. 2018.

[2] T. Yang, Z. Chaoqiang, W. Jianrui. "Perception and navigation in autonomous systems in the era of learning: A survey". IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no 12, p. 9604-9624. 2022.

[3] A. Sara, I. Halima, K. Ali. "Advancing autonomous vehicle control systems: An in-depth overview of decision-making and manoeuvre execution state of the art". The Journal of Engineering, vol. 223, no 11, p. e12333. 2023.

[4] S. Afroosheh, and M. Askari. "Geospatial Data Fusion: Combining Lidar, SAR, and Optical Imagery with AI for Enhanced Urban Mapping". arXiv preprint. 2024.

[5] S. Du, W. Li, and F. Ling. "GeoAI in terrain analysis: Enabling multi-source deep learning and data fusion for natural feature detection". Environmental Modelling & Software, 143, 105103. 2021.

[6] LR, Rolen, BHAT, Soumya J., and KV, Santhosh. "Prospective study on challenges faced in a perception system". Cogent Engineering, vol. 11, no 1, p. 2353498. 2024.

[7] JIN, Ronghe, ZHANG, Guohao, HSU, Li-Ta. "A survey on cooperative positioning using GNSS measurements". IEEE Transactions on Intelligent Vehicles, 2024.

[8] LI, Xiaoyu, GUO, Xiye, LIU, Kai, et al. "Context Awareness Assisted Integration System for Land Vehicles". Electronics, vol. 13, no 11, p. 2038. 2024.

[9] Kadrolli, V., and Kalnoor, G. "AI-Infused Strategies for Mitigating Uncertainty in Continental-Scale Surface Mass Change Analysis Through GPS and GRACE-FO". Remote Sensing in Earth Systems Sciences, 7(1), 80–89. 2024.

[10] Li, W., and Wang, S. "GeoAI: Where machine learning and big data converge in GIScience". Journal of Spatial Information Science, vol. 20, 71–83. 2020.

[11] Liu, Y., and Fan, H. "Multi-sensor integrated navigation/positioning systems using data fusion: From analytics-based to learning-based approaches". Information Fusion, 95, 62–90. 2023.

[12] Niantic Inc. "Niantic is building a 'geospatial' AI model based on Pokémon Go player data". The Verge. 2024.

[13] SandboxAQ. "The Startup That Wants to Revolutionize Satellite Navigation". Wired. 2024.

[14] Sharma, A., and Gupta, R. "AI-driven Real-time System for Land Surveillance and Reconnaissance". Proceedings of the 2022 Fourteenth International Conference on Contemporary Computing (IC3-2022), 710–715.

[15] Wang, S., and Li, W. "Automated terrain feature identification from remote sensing imagery: A deep learning approach". International Journal of Geographical Information Science, 35(4), 703–726. 2021.

[16] Zhang, T., and Chen, L. "Deep learning-based approach for landform classification from integrated data sources of digital elevation model and imagery". Geomorphology, 357, 107091. 2020.

[17] K. Zhu and T. Zhang. "Deep reinforcement learning based mobile robot navigation: A review". Tsinghua Science and Technology, vol. 26, no. 5, pp. 674–691, 2021.

[18] S. Bijjahalli, R. Sabatini, and A. Gardi. "Advances in intelligent and autonomous navigation systems for small uas". Progress in Aerospace Sciences, vol. 115, p. 100617, 2020.

[19] J. Crespo, J. C. Castillo, O. M. Mozos, and R. Barber. "Semantic information for robot navigation: A survey". Applied Sciences, vol. 10, no. 2, p. 497, 2020.

[20] S. Rezwan and W. Choi. "Artificial intelligence approaches for uav navigation: Recent advances and future challenges". IEEE access, vol. 10, pp. 26 320–26 339, 2022.

[21] N. F. A. M. Fadzil, H. M. Fadzir, H. Mansor, and U. Rahardja. "Driver behaviour classification: A research using obd-ii data and machine learning". Journal of Advanced Research in Applied Sciences and Engineering Technology, pp. 51–61, 2024.

[22] S. Jordan and T. M. Nguyen. "Machine learning in autonomous vehicles: Current status and future directions". IEEE Transactions on Vehicular Technology, vol. 72, no. 6, pp. 7113–7128, 2023.

[23] Sierra N. Young. "Intelligent robots for agriculture - Ag-robot development, navigation, and information perception". Frontiers in Robotics and AI. Section Industrial Robotics and Automation. Volume 12 – 2025.

[24] Milad Rahmat. "Edge AI-Powered Real-Time Decision-Making for Autonomous Vehicles in Adverse Weather Conditions". ArXiv, March 2025.

[25] Shumaila Javaid, Muhammad Asghar Khan, and Hamza Fahim. "Explainable AI and monocular vision for enhanced UAV navigation in smart cities: prospects and challenges". Frontiers in Sustainable Cities. Volume 7 – March 2025.

[26] Na Tang, Yuehui Liao, Yu Chen, and Guang Yang. "An AI-Driven Vision Sensor Framework for High-Precision, Real-Time Video Portrait Segmentation with Enhanced Temporal Consistency and Optimized Model Design". Journal Sensors. 25(5). 2025.

[27] G C Sunil, Arjun Upadhyay, and Xin Sun. "Development of software interface for AI-driven weed control in robotic vehicles, with time-based evaluation in indoor and field settings". Smart Agricultural Technology-Elsevier. Volume 9, December 2024.

[28] Mohsen Soori, Behrooz Arezoo, and Roza Dastres . "Artificial intelligence, machine learning and deep learning in advanced robotics, a review". Cognitive Robotics. Volume 3. Pages 54-70. 2023.

[29] Christos Gkrizis, Nikos Dimitropoulos, Konstantinos Katsampiris-Salgado. "Intersubjective AI-driven multimodal interaction for advanced user-centric HRC applications - the JARVIS approach". Procedia CIRP-Elsevier. Volume 130. Pages 325-330. 2024.

[30] Omsri Aeddula, Martin Frank, and Ryan Ruvald. "AI-Driven Predictive Maintenance for Autonomous Vehicles for Product-Service System Development". Procedia CIRP- Elsevier. Volume 128. Pages 84-89. 2024.

SEARCHX: An Integrated Framework of Distributed Intelligent Search Services Based on Web Browser

Zehui Zhang*, Lin Zhou, Jie Peng, Liwei Wang, Bo Cheng
Inner Mongolia Power Digital Research Institute, Hohhot 010000, China

***Abstract*—To address the growing demand for web search and improve the performance and accuracy of search systems, this study proposes a distributed intelligent search service integration framework based on SEARCHX. This framework leverages the local computational power of the browser, integrating inverted indexing, data sharding, and replication mechanisms, as well as the Term Frequency-Inverse Document Frequency (TF-IDF) intelligent ranking algorithm. These components enable front-end distributed processing of search tasks and multi-source result fusion. Experiments are conducted on six major browser platforms, Chrome, Firefox, Edge, Safari, etc., using the open-source Text REtrieval Conference (TREC) dataset. The system's response performance and accuracy are evaluated under varying search loads. The experimental results show that, compared to the unoptimized version, the optimized SEARCHX reduces the average response time by approximately 27 per cent under medium-to-high load conditions. Precision improves by an average of 0.05, and the F1 score increased by more than 0.04 on all platforms. The system also demonstrates good stability and consistency across multiple platforms. SEARCHX provides a viable approach to building decentralized, high-efficiency, and easily deployable intelligent search services, with strong practical value and expansion potential. This study aims to construct a decentralized, cross-platform, and high-performance intelligent search service framework, offering a more efficient, stable, and accurate technical support solution for users in complex search environments.**

Keywords*—Web; TF-IDF; distributed network; intelligent search; SEARCHX

I. Introduction

In the context of rapid advancements in information technology, internet data is experiencing exponential growth, and users face unprecedented challenges of information overload and retrieval difficulties in the web environment. As the core tool for information access, intelligent search engines have permeated multiple fields such as knowledge acquisition, data mining, and personalized recommendation systems [1]. However, mainstream search engines are predominantly based on centralized server architectures. While they offer powerful computational capabilities and global indexing services, they increasingly show issues in handling massive concurrent requests, ensuring user privacy, and supporting flexible deployments. These problems include high latency, poor scalability, high deployment costs, and data security risks [2]. Meanwhile, web browser technology is undergoing significant evolution. New-generation browser-side technologies, represented by JavaScript engine optimizations, WebAssembly (Wasm), and Service Workers, have transformed web browsers from mere information display tools into powerful computing platforms [3]. The enhanced local storage, local execution, and multi-threading capabilities of browsers provide a practical foundation for running high-performance applications directly within the browser. This shift enables parts or even the entire search process to be offloaded to the browser, which alleviates server load, reduces response latency, enhances data localization capabilities, and strengthens user privacy protection [4]. In enterprise-level information systems, distributed search architectures like Elasticsearch are widely deployed, relying on mechanisms such as inverted indexing, data sharding, and replica redundancy to improve search efficiency and system fault tolerance [5]. However, these technologies are still primarily server-side and have not been effectively migrated or integrated into the browser side, leaving a gap in research and practical application [6]. Therefore, fully leveraging the potential of browsers as client-side computing platforms to construct a cross-platform, intelligent, distributed search framework becomes an important and worthy direction for further exploration.

Based on the above background, this study proposes the SEARCHX framework, a distributed search engine optimization solution that integrates inverted indexing, data sharding, replica synchronization, and the Term Frequency–Inverse Document Frequency (TF-IDF) intelligent ranking algorithm. The framework is introduced within the browser environment to enable distributed task scheduling, plugin-based service integration, and local search result fusion. This design supports cross-platform deployment and flexible scalability, and significantly enhances system response efficiency and user data privacy control capabilities. The main innovation of this study is to systematically migrate the complete distributed search engine architecture—including inverted index, data sharding, replica synchronization, and TF-IDF intelligent ranking algorithm—to the Web browser side. This design is expected to realize truly decentralized, cross-platform search processing and local multi-source result fusion, thus filling the gap in the research and practice of a complete distributed search architecture on the browser side.

The subsequent structure of this study is arranged as follows: Section II reviews the related work on intelligent search, distributed architecture, and browser computing. Section III elaborates on the design of the SEARCHX framework and distributed optimization methods in detail. Section IV presents the performance evaluation and comparative experimental results under multi-browser platforms. Section V summarizes the research results and looks forward to future directions. Through systematic discussion and empirical analysis, this study aims to comprehensively present the technical implementation

*Corresponding author.

and performance advantages of the SEARCHX framework to readers.

II. Related Work

A. Distributed Intelligent Search Algorithm

In recent years, both academia and industry have conducted extensive research on intelligent search systems, distributed architectures, and front-end computing, resulting in several valuable development paths [7]. In the field of intelligent search, Wang et al. discovered that pre-trained encoders such as Bidirectional Encoder Representations from Transformers (BERT) were mainstream methods for information retrieval. BERT could be categorized into six technical paths, including long document processing, semantic integration, and efficiency balancing. Their results indicated that BERT-based models outperformed traditional methods in terms of accuracy and adaptability, with lower computational costs [8]. He et al. proposed that personalized recommendation mechanisms could enhance search user satisfaction and result relevance by modeling user historical behaviors and interest preferences [9]. In the domain of distributed search systems, Fan et al. found that inverted index structures significantly improved recall and precision in large-scale text retrieval, making it a core technology for building high-performance search engines [10]. Merlin and Prem identified that the MapReduce framework based on the Jaya-Sine Cosine Algorithm (Jaya-SCA) enabled efficient indexing and retrieval in big data. This method achieved an F1 score of 0.5323, a recall rate of 0.4400, and a precision rate of 0.6867 on the StatLog heart disease dataset, significantly improving information retrieval performance [11]. Soltanmohammadi et al. found that a consistent hashing-based data partitioning strategy effectively avoided data loss due to node failures [12].

B. Browser-Side Computing and Search Results Sorting

Regarding browser-side computing and collaboration mechanisms, Kjorveziroski and Filiposka highlighted that, with the development of WebAssembly, modern browsers had the ability to execute local computational tasks, showing potential to serve as edge nodes in distributed systems [13]. Putra et al. designed the SearchX platform, which supported user behavior tracking and collaborative task allocation, demonstrating that browsers could serve as both execution and collection terminals for search services. However, their system mainly focused on research purposes and lacked complete distributed indexing and search capabilities [14]. In ranking algorithms, Dai et al. found that while the TF-IDF algorithm was simple, it effectively reflected keyword importance and document relevance in static document collections, serving as the foundation for search ranking [15]. Yang and Choi discovered that the Best Matching (BM) 25 algorithm, by introducing a document length normalization factor, further improved ranking accuracy [16]. In search service integration systems, Dejonckheere found that search service platforms based on microservice architectures effectively enhanced module decoupling and deployment flexibility, widely applied in cloud computing and large-scale web services [17].

In summary, although significant progress has been made in areas such as distributed search, intelligent ranking, and browser-side computing, there is currently a lack of a unified integrated platform that can simultaneously support local browser computation, cross-node collaborative search, and intelligent result fusion. The SEARCHX framework proposed here builds upon the aforementioned research achievements and fills the gap in browser-side distributed search integration. This framework holds significant theoretical and practical value.

III. Research Methodology

A. SEARCHX Framework

SEARCHX is a distributed intelligent search service integration framework based on modern web browsers, designed to leverage the computational and communication capabilities of browsers to provide a low-barrier, high-performance collaborative search experience. Its architecture adopts a client-server model, with the front-end implemented using the React framework for modular and component-based design. It supports multi-browser cross-platform access without the need for additional software installations. Users can simply participate in search tasks by specifying a Uniform Resource Locator (URL). The front-end is responsible for displaying the user interface, managing search sessions, and collecting user operation logs in real-time. The back-end, based on the Node.js platform, uses Express and Socket.io to handle efficient Hypertext Transfer Protocol (HTTP) requests and WebSocket real-time communication, managing user grouping, task synchronization, and data storage [18]. For database management, SEARCHX utilizes MongoDB to support dynamic data structures, enabling flexible storage of massive logs and experimental data. The basic framework of SEARCHX is illustrated in Fig. 1.

B. SEARCHX Framework Based on Distributed Search Optimization

To meet the high-performance requirements of massive data searches and enhance the accuracy and stability of the system, the SEARCHX framework introduces several distributed search optimization techniques on the browser side. These techniques include the construction of inverted indexes, the TF-IDF ranking algorithm, data sharding and replication mechanisms, as well as task scheduling and result fusion based on real-time communication.

Inverted indexing is the core data structure of modern search engines. It establishes a mapping from keywords to document lists, enabling fast keyword location and retrieval. In a distributed environment, local construction of inverted indexes helps improve search speed and the system's parallel processing capabilities [19]. Specifically, the inverted index is composed of a keyword and the inverted list of documents. The calculation equation is as follows:

$$I = \{ (t_i, L_i) \mid t_i \in T \} \tag{1}$$

$$L_i = \{ (d_j, f_{ij}) \mid d_j \in D, f_{ij} > 0 \} \tag{2}$$

I is an inverted index set. t_i is the i-th keyword in the keyword set T. L_i is the inverted list corresponding to the keyword t_i. d_j is the j th document in the document set D. f_{ij} is the word frequency of keyword t_i in document d_j.

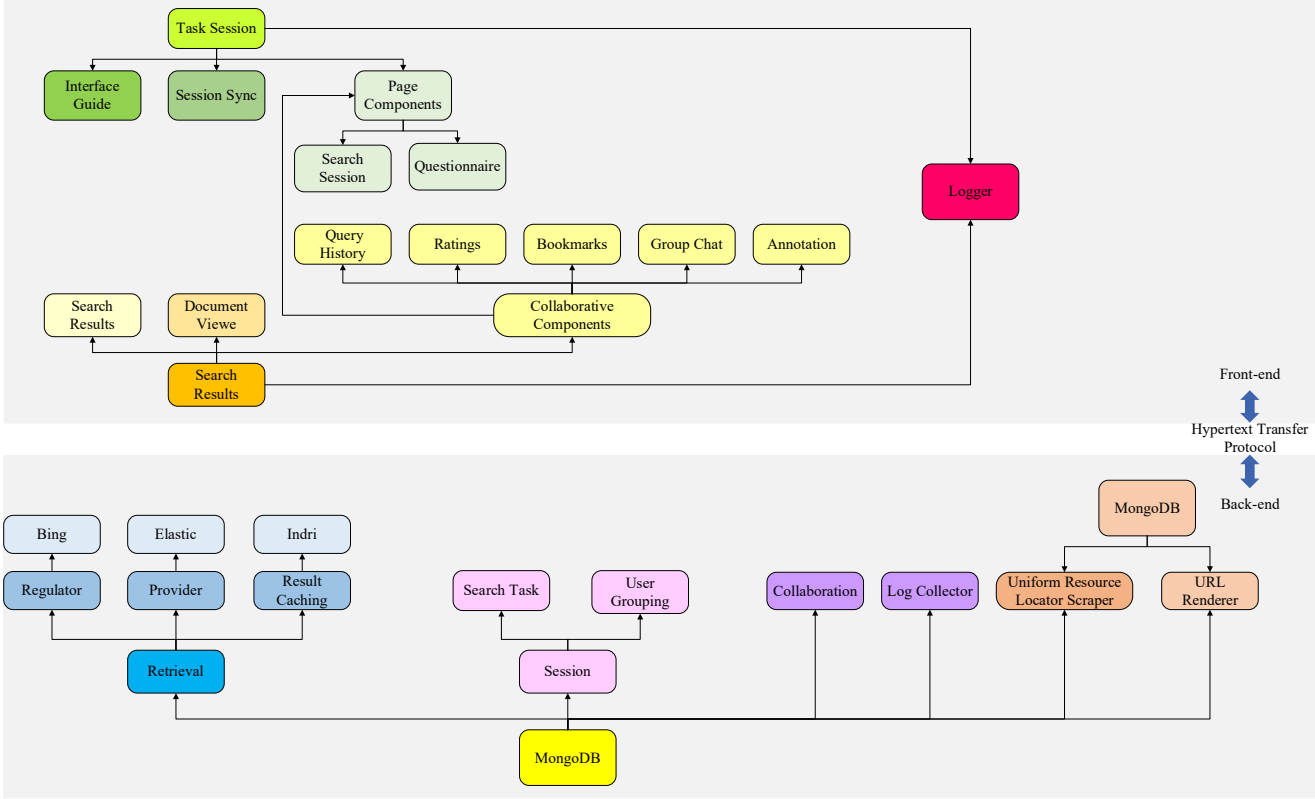


Fig. 1. SEARCHX basic framework.

The TF-IDF algorithm is a classical weighting method based on the vector space model, which combines the frequency of keywords in a single document and their rarity in the whole document set, thus weighting keywords and highlighting important and highly differentiated words [20-22]. The specific calculation equation is as follows:

$$tf_{ij} = \frac{f_{ij}}{\max_{t_k \in d_j} f_{kj}} \tag{3}$$

$$idf_i = \log\left(\frac{N}{df_i + 1}\right) \tag{4}$$

$$TF - IDF(t_i, d_j) = tf_{ij} \cdot idf_i \tag{5}$$

tf_{ij} is the normalized word frequency. idf_i is the inverse document frequency. df_i is the number of documents containing the keyword t_i. N is the total number of documents.

The vector space model represents documents and queries as multidimensional vectors, and quantifies the correlation between them by calculating the cosine similarity of the included angle between the vectors [23]. This measurement method not only overcomes the rigid limitation of the traditional Boolean model but also flexibly reflects the similarity between texts and improves the accuracy of search matching. The calculation equation of the vector space model is as follows:

$$\mathbf{v}_{d_j} = (v_{j1}, \ldots, v_{jM}) \tag{6}$$

$$v_{ji} = TF - IDF(t_i, d_j) \tag{7}$$

$$\mathbf{v}_Q = (w_1, \ldots, w_M) \tag{8}$$

$$w_i = \begin{cases} w_i, & t_i \in Q \\ 0, & otherwise \end{cases} \tag{9}$$

$$Sim(d_j, Q) = \frac{\mathbf{v}_{d_j} \cdot \mathbf{v}_Q}{\|\mathbf{v}_{d_j}\| \cdot \|\mathbf{v}_Q\|} \tag{10}$$

$\mathbf{v}_{d_j}$ is the vector representation of document d_j. $\mathbf{v}_Q$ is the vector representation of query Q. M is the total number of keywords. w_i is the weight of keyword t_i in the query. $Sim(d_j, Q)$ is the similarity value between the document and query.

Sharding is a commonly used data partitioning method in distributed systems, where large-scale data is horizontally split across multiple nodes to reduce the load on individual nodes and enable parallel processing. A hash function ensures that data is evenly distributed, preventing load imbalance across nodes [24]. The sharding mechanism is a key technology for achieving system scalability and handling high concurrency. The specific partitioning expression is as follows:

$$S_m = \{d_j \in D \mid h(d_j) mod K = m\} \tag{11}$$

S_m is the m-th slice. $h(d_j)$ is the hash value of document d_j. K is the total number of slices. m is the slice number.

The replication mechanism enhances the fault tolerance and availability of the system by storing identical data across multiple nodes. In the event of a node failure, the system can switch to a backup replica, ensuring continuous and stable

operation [25]. Additionally, replication helps with load balancing and improves access efficiency. The replica selection algorithm can be expressed as follows:

$$R_m = \{n_{m1}, n_{m2}, \ldots, n_{mR}\} \tag{12}$$

$$n^* = arg\min_{n \in R_m} Load(n) \tag{13}$$

R_m is the copy set of the m-th slice. n_{m1} is the replica node. $Load(n)$ is the current load value of node n. n^* is the selected replica node.

Parallel computing divides large tasks into smaller sub-tasks and assigns them to different computation nodes for simultaneous execution, reducing the overall response time. Search tasks are split into shard queries, and high-efficiency collaboration between nodes is achieved through modern Web asynchronous communication protocols [26]. The system response time can be calculated as follows:

$$T_{response} = \max_{0 \le m < K} T_m \tag{14}$$

$T_{response}$ is the overall response time of the system. T_m is the query response time of the m-th fragment.

Multi-source result fusion, based on data fusion theory, integrates partial results from each shard, eliminates duplicates, and forms the final ordered result set [27]. Merge sorting ensures the correctness of the overall ranking, while deduplication prevents result redundancy, enhancing the user experience. The specific calculation equation is as follows:

$$R = \bigcup_{m=0}^{K-1} R_m \tag{15}$$

$$R_{final} = \delta\varepsilon\upsilon\pi\lambda\iota\chi\alpha\tau\varepsilon(R) \tag{16}$$

R is the combination result set. R_{final} is the result set after final sorting and deduplication.

The load balancing theory provides a dynamic adjustment strategy for task allocation. By monitoring the node load in real-time and combining it with the weighted smoothing algorithm, the task scheduling weight is adjusted to prevent the overload of a single node and improve the system stability and resource utilization [28]. The dynamic updating equation of scheduling weight is as follows:

$$w_n^{(t+1)} = \alpha w_n^{(t)} + (1-\alpha)\frac{1}{Load(n)} \tag{17}$$

The maximum concurrency control calculation equation is as follows:

$$C = min\left(\frac{X}{Tt}, C_{max}\right) \tag{18}$$

$w_n^{(t)}$ is the scheduling weight of node n at time t. α is a smoothing factor (0~1). X is the total number of tasks to be processed. Tt is the time window size. C_{max} is the maximum number of concurrent tasks supported by the system.

The process of SEARCHX based on distributed search optimization is shown in Fig. 2.

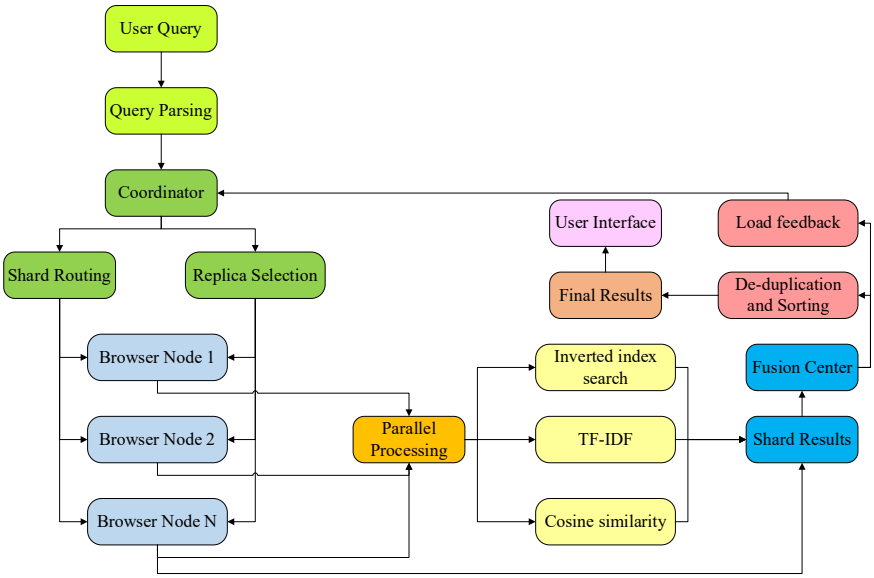


Fig. 2. The process of SEARCHX based on distributed search optimization.

C. System Assembly Design

The experimental platform of this study consists of six identical physical computers, each equipped with an Intel Core i7-10700 CPU, 6GB of RAM, and 512GB solid-state drive storage. The operating system on all machines is Linux Ubuntu 20.04 Long-term Support (LTS). The latest versions of mainstream web browsers, including Chrome, Firefox, Edge, Safari, Opera, and Brave, are deployed on each node to test the compatibility and performance differences of SEARCHX in a multi-platform browser environment. The software development environment includes Node.js v18.0 for the backend service runtime, React v18.2 for frontend component development, and MongoDB v6.0 for lightweight local indexing and data storage. Additionally, the system communication layer incorporates WebSocket and Web Real-Time Communications (WebRTC) protocols to support distributed collaborative search across browsers.

SEARCHX adopts a modular and component-based system deployment strategy. The main node runs the backend service responsible for task scheduling and user sessions, while the browser nodes act as independent search sub-nodes that load the frontend interface and locally execute indexing and query tasks. The deployment process includes the following steps: 1) Configure Node.js and Express backend services on the main node and open the WebSocket communication port. 2) Run the frontend page on each browser node, loading the SEARCHX interface via a unified URL. 3) Each browser node autonomously registers and identifies its role based on the system initialization logic. 4) Use browser-side Service Workers and localStorage to achieve local data persistence and offline accessibility. This architecture allows the browser to function both as a lightweight client for displaying the interface and as a search computation node, enabling a truly decentralized search system at the edge.

The testing plan is divided into two stages: functional testing and performance testing. In functional testing, the system verifies the functionality of critical modules, including: task scheduling, index construction, result sorting, node communication, and failover mechanisms. The goal is to ensure that these modules are operational and functioning as expected within the system. In performance testing, the system is tested under different search request volumes (10, 100, 500, 1000, 2000, 5000 requests). The tests are conducted on six major browser platforms, and the system's response time, stability, and other performance metrics are evaluated. Each node automatically logs data, which is sent back to the main node for analysis and assessment.

This study uses the Text REtrieval Conference (TREC) open text retrieval dataset as the training and evaluation data source for the system [29, 30]. This study selects this dataset mainly based on the following considerations: the dataset includes 250 real user queries, more than 520,000 news documents, and manually annotated relevance judgments. It features diverse query topics, moderate document scale, and high annotation quality, which can comprehensively evaluate the precision, recall, and system stability of the search engine under different query complexities and document scales, and meet the needs of this study for empirical verification of retrieval performance. The subset used in this study is the TREC 2004 Robust Track, which includes 250 standard query tasks, 528,000 English news and article documents, and manually annotated relevance grade labels. This makes it suitable for analyzing precision, recall, accuracy, and F1 score metrics.

IV. Experimental Evaluation of SEARCHX Performance

A. Results of Performance Analysis

The comparison results of the performance changes of the search framework under different platforms with the query volume are shown in Fig. 3 to Fig. 8.

Analysis of Fig. 3 shows that, as the number of query entries increases, the average response time for all browsers tends to rise, indicating that an increase in search request volume has a significant impact on system response. Among all browsers, Chrome performs the best, with the shortest response time, suggesting that its JavaScript engine and network communication optimizations are more suitable for distributed search loads. Safari, on the other hand, performs relatively poorly, with the longest response time. The increase in response time accelerates as the query volume grows, and for 5000 queries, the average response time is approximately 6 to 7 times higher than that for 10 queries. It indicates that the system can still maintain relatively stable response performance even under large-scale tasks.

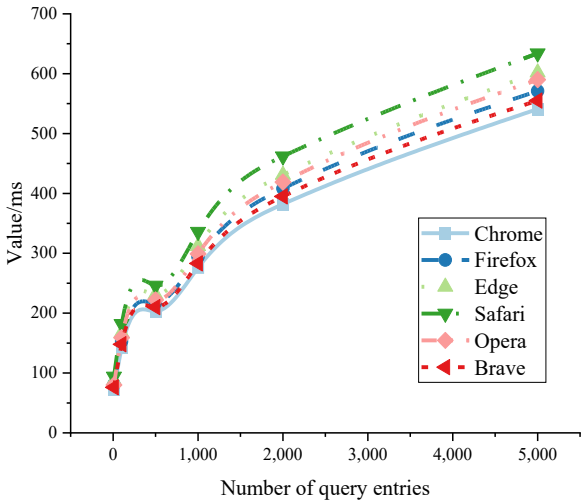


Fig. 3. Comparison results of average response time.

Fig. 4 shows that the CPU utilization increases significantly with the growth of query volume, reflecting the increasing computational resource consumption of search tasks. Chrome and Brave demonstrate lower CPU usage under varying query loads compared to other browsers, indicating more efficient resource management. Safari and Edge exhibit the highest CPU utilization under heavy loads, which could potentially affect their ability to handle multiple tasks concurrently. Overall, in distributed tasks, it is essential to allocate the load reasonably in this system to avoid CPU bottlenecks on a single node.

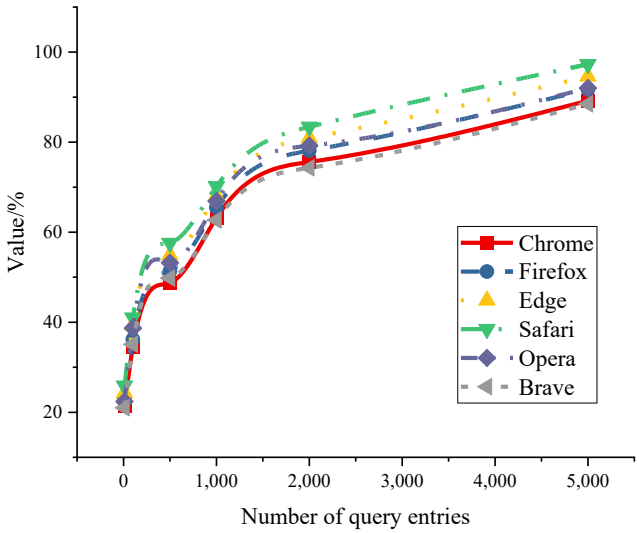


Fig. 4. CPU occupancy comparison results.

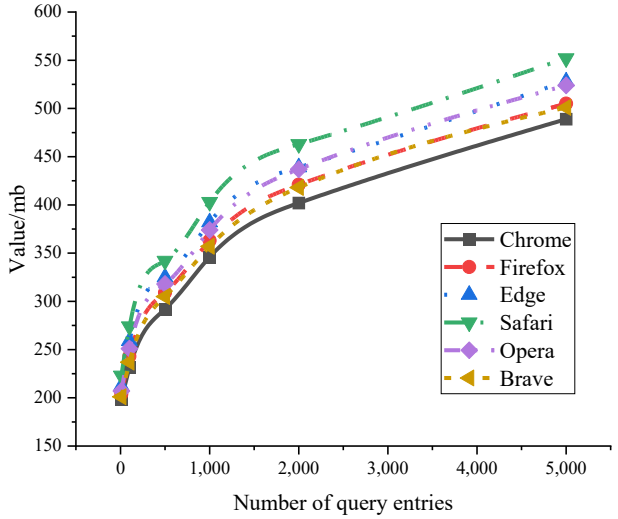


Fig. 5. Memory usage comparison result.

Fig. 5 reveals that memory consumption in all browsers increases linearly with the number of query entries, indicating that search result indexing and cached data consume a significant amount of memory. Chrome and Brave exhibit relatively lower memory usage, suggesting optimized memory management and recycling mechanisms. Safari, on the other hand, shows the highest memory consumption, indicating potential issues with memory leakage or room for improvement in cache management strategies.

Fig. 6 shows that the throughput rate decreases as the query volume increases on all platforms, indicating a certain processing capacity bottleneck. Chrome maintains the highest throughput, handling more requests consistently, which reflects its strong support for distributed search tasks. Edge and Safari, however, exhibit lower throughput, likely due to limitations in network communication efficiency and multi-threading processing capabilities. This result underscores the critical importance of improving network communication and concurrent processing technologies for system performance optimization

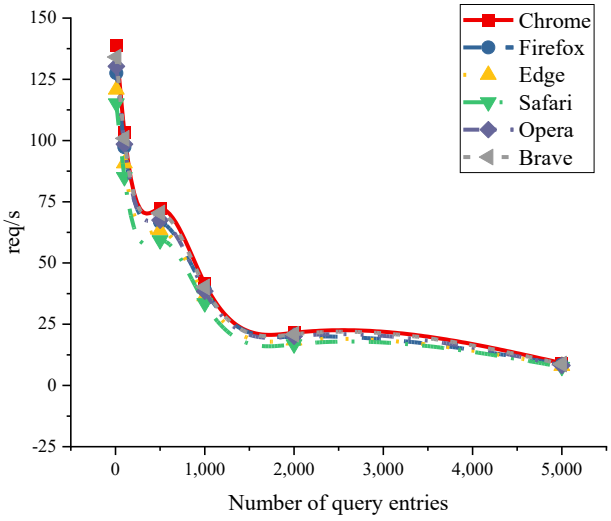


Fig. 6. Throughput comparison results.

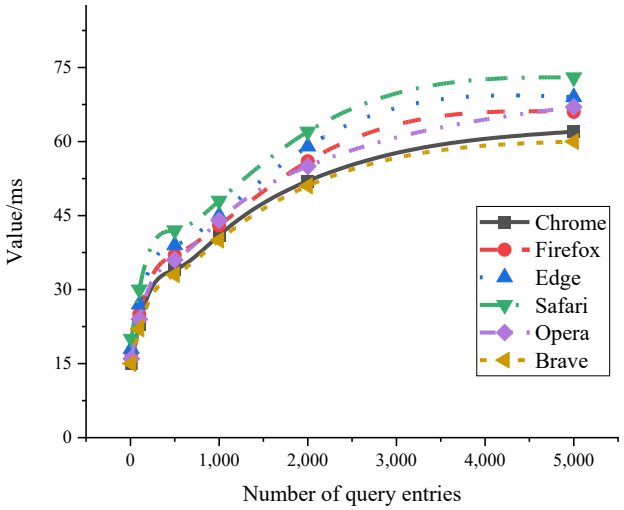


Fig. 7. Data synchronization delay result.

Fig. 7 shows that data synchronization latency increases as the query volume grows, indicating a higher communication burden between nodes. Chrome and Brave maintain the lowest synchronization delay, reflecting superior WebSocket or WebRTC communication performance. Safari, on the other hand, exhibits the highest latency, which could degrade the user experience, particularly in collaborative search scenarios. The study suggests that optimizing the synchronization mechanism is a key factor in enhancing system response speed and stability.

Fig. 8 reveals that the overall score decreases as the query load increases, reflecting the system's stability pressure under high load conditions. Chrome and Brave maintain higher scores than other platforms, indicating better stability in complex load environments. Edge and Safari have lower scores. It suggests that future optimization of the SEARCHX framework should focus on improving error recovery and resource management strategies to ensure stable operation during prolonged high-load conditions.

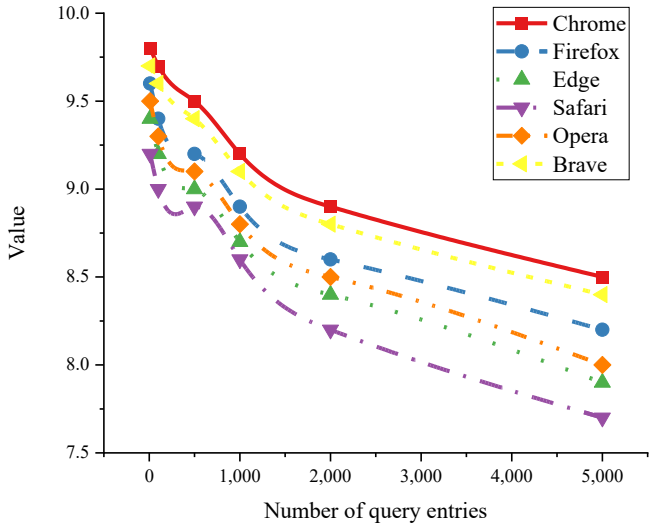


Fig. 8. System stability score results.

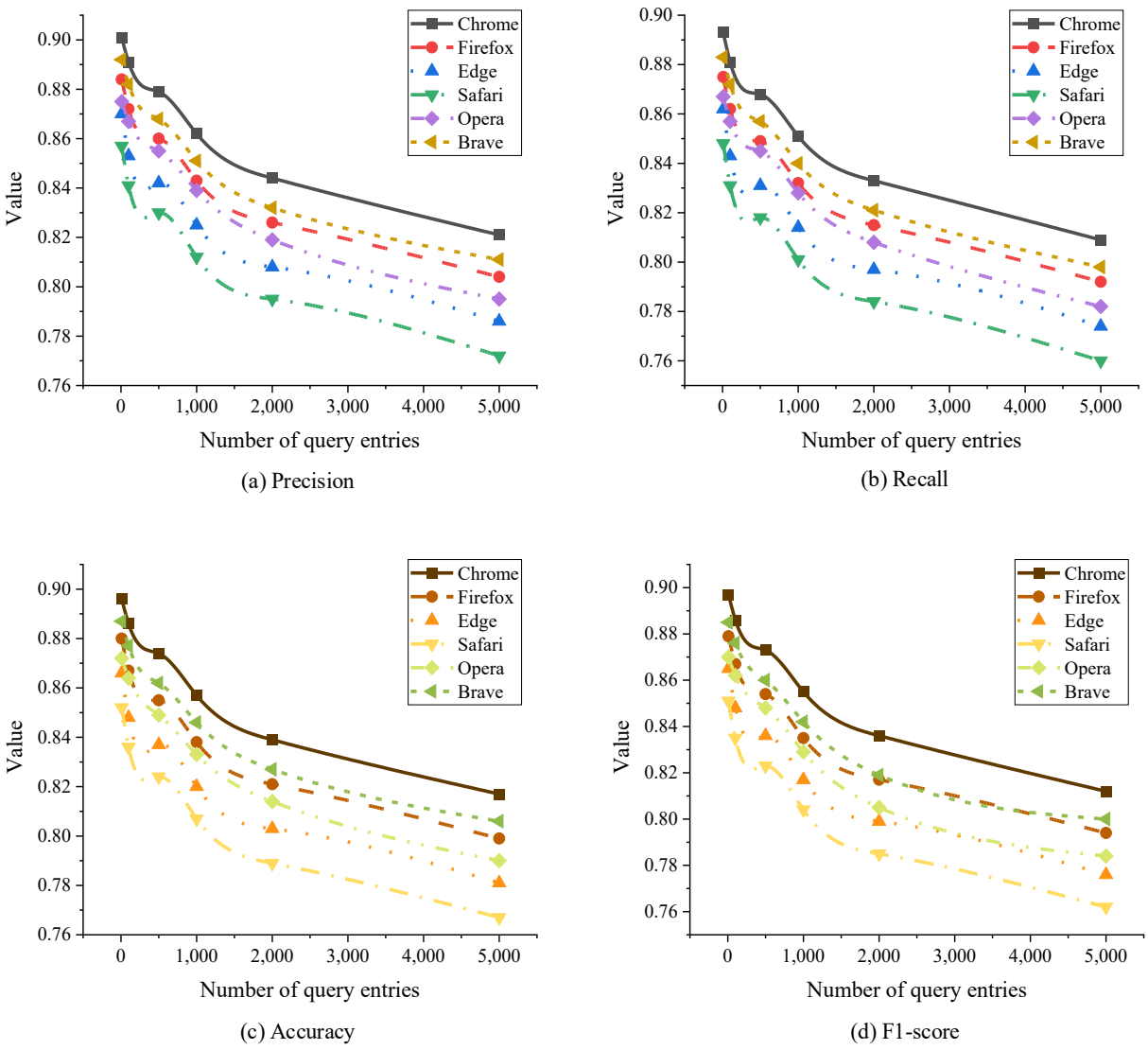


Fig. 9. Comparison of query accuracy under different platforms.

The comparison of query accuracy under different platforms as the query volume increases is shown in Fig. 9.

Fig. 9 shows that as the number of query entries increases from 10 to 5000, the search result accuracy across all browser platforms exhibits a gradual decline. Taking Chrome as an example, the accuracy drops from 0.901 to 0.821, but it remains at a high level, indicating excellent performance in intelligent ranking and index optimization. Firefox and Brave follow closely, with accuracy dropping from 0.884 and 0.892 to 0.804 and 0.811, respectively. Safari and Edge show slightly weaker performance, starting at 0.857 and 0.870, dropping to 0.772 and 0.786, but still maintaining stability. Regarding recall rate, Chrome and Brave platforms decrease from 0.893 and 0.883 to 0.809 and 0.798, demonstrating strong document coverage and distributed indexing effects. Firefox and Opera decrease from 0.875 and 0.867 to 0.792 and 0.782. Safari and Edge platforms start lower, at 0.848 and 0.862, and drop to 0.760 and 0.774, indicating room for improvement in shard synchronization and index coverage for these two platforms. As the query load increases, the accuracy of all platforms gradually decreases but remains stable. Chrome's accuracy drops from 0.896 to 0.817, and Brave's from 0.887 to 0.806, showing the best performance. Firefox and Opera start at 0.880 and 0.872, and end at 0.799 and 0.790, respectively. Safari and Edge show slightly lower accuracy, from 0.852 and 0.866 to 0.767 and 0.781. The F1 scores for all platforms decline as the number of queries increases, but the minimum remains above 0.760. Chrome and Brave show the best overall performance, with F1 scores dropping from 0.897 and 0.885 to 0.812 and 0.800, respectively. Firefox and Opera's F1 scores drop from 0.879 and 0.870 to 0.794 and 0.784, still maintaining a high level. Safari and Edge's F1 scores decrease from 0.851 and 0.865 to 0.762 and 0.776, slightly lower than the others but still stable overall.

The results of performance comparison between the search framework and the unoptimized framework under different platforms are shown in Table I.

Table I shows that the optimized SEARCHX search framework outperforms the unoptimized version across six major browser platforms, demonstrating higher system

performance and search quality. Specifically, the average response time decreases by approximately 27 per cent, while CPU and memory usage significantly decrease, indicating improved efficiency in resource scheduling and task management. Throughput increased by an average of 34.8 req/s, greatly enhancing the system's ability to handle concurrent requests. In terms of accuracy, precision and F1 scores improve by an average of approximately 0.05. The Safari platform shows the most significant improvement, with a precision increase of 0.064, highlighting the stability and practicality of the optimized framework in large-scale distributed query environments.

TABLE I. THE PERFORMANCE COMPARISON RESULTS OF THE SEARCH FRAMEWORK AND UNOPTIMIZED FRAMEWORK UNDER DIFFERENT PLATFORMS IN THIS STUDY

Platform	Framework Version	Avg. Response Time (ms)	CPU Usage (%)	Memory Usage (mb)	Throughput (req/s)	Precision	F1 Score
Chrome	Unoptimized	328	72.4	946	143.7	0.762	0.769
	The proposed model	238	63.1	802	178.5	0.821	0.812
Firefox	Unoptimized	349	74.8	971	139.1	0.743	0.751
	The proposed model	257	65.6	835	168.3	0.804	0.794
Edge	Unoptimized	367	77.2	990	135.2	0.725	0.731
	The proposed model	274	68.5	857	162.4	0.786	0.776
Safari	Unoptimized	382	79.1	1014	129.5	0.708	0.717
	The proposed model	293	70.2	876	157.2	0.772	0.762
Opera	Unoptimized	341	73.6	938	142.1	0.735	0.742
	The proposed model	249	64.9	819	171.3	0.795	0.784
Brave	Unoptimized	330	71.9	925	145.5	0.754	0.761
	The proposed model	241	62.8	798	177.8	0.811	0.8

B. Discussions

This study achieves significant improvements over previous research in multiple dimensions, including performance, system resource utilization, and search quality, by constructing and optimizing the SEARCHX search framework. Unlike previous models that applied BERT for information retrieval, which excels in semantic understanding but is limited by computational resources, this study focuses on a lightweight search architecture with strong deployability. By employing TF-IDF and distributed inverted indexing, the framework reduces average response time, lowers CPU utilization, and optimizes memory consumption across six major browser platforms, while also improving average throughput. In contrast to previous approaches that used optimization algorithms to enhance recall performance, this study shows an average improvement of approximately 0.043 in F1 score and more than 0.05 in precision, demonstrating superior practicality and platform adaptability. Additionally, building on previous browser collaboration models that are limited to laboratory settings, SEARCHX has successfully implemented a real-world framework supporting local computation and multi-node collaboration. In summary, this study inherits the theoretical advancements in search quality enhancement, system fault tolerance, and architecture design from prior literature. It also makes a substantial breakthrough in performance, efficiency, and adaptability by constructing a unified integrated platform and conducting cross-platform experimental validation. This study fills the gap in existing search systems related to front-end collaboration and resource optimization, providing a new paradigm for the edge deployment and platform independence of future intelligent search systems.

V. CONCLUSION

This study systematically evaluates the performance of the optimized SEARCHX search framework across six major browser platforms, demonstrating significant improvements in several key metrics. In terms of performance, the optimized system reduces the average response time from approximately 350ms to 250ms, representing a 27 per cent decrease. CPU utilization drops by nearly 9 per cent (for instance, Chrome went from 72.4 per cent to 63.1 per cent). Memory consumption decreases by an average of over 140MB, reflecting more efficient system resource management. Regarding throughput, the average increase is about 34.8 req/s, with Chrome showing the most significant improvement, from 143.7 req/s to 178.5 req/s, greatly enhancing concurrent processing capabilities. In terms of search quality, the optimized SEARCHX framework achieves an average improvement of over 0.05 in precision, with Safari showing the most substantial increase, from 0.708 to 0.772, a rise of 0.064. The F1 score also improves, with an average increase of 0.04-0.05, effectively balancing system recall and precision. In summary, the optimized SEARCHX framework outperforms in four key areas: response speed, system resource usage, throughput performance, and search quality. The results demonstrate its strong stability, high efficiency, and high accuracy in multi-platform, large-scale query environments, making it highly promising for broad applications. This study selects this dataset mainly based on the following considerations: the dataset includes 250 real user queries, more than 520,000 news documents, and manually annotated relevance judgments. It features diverse query topics, moderate document scale, and high annotation quality, which can comprehensively evaluate the precision, recall, and system stability of the search engine under different query complexities

and document scales, and meet the needs of this study for empirical verification of retrieval performance.

Although this study verifies the excellent performance of the SEARCHX framework in a laboratory environment, its deployment in real network environments still faces many challenges. For example, heterogeneous network conditions and differentiated browser computing capabilities may affect the collaboration efficiency and result consistency among nodes; although local data processing enhances user privacy protection, it also brings new client-side security considerations; in addition, when facing Internet-scale ultra-large datasets, the current architecture still needs further optimization in terms of index construction and update efficiency. Identifying these limitations is crucial for the future development and practical application of the framework.

Based on this, future research will conduct an in-depth exploration along the following directions: 1) Attempt to introduce a dynamic load balancing algorithm to further improve the collaboration efficiency among heterogeneous browser nodes. 2) Explore the integration of lightweight semantic models with the existing TF-IDF algorithm to enhance the ability to understand complex semantic queries. 3) Extend the framework to the mobile browser environment and evaluate its performance, focusing on optimizing the index efficiency under large-scale data, and simultaneously explore the integration of lightweight neural ranking models to improve the ability to understand complex queries.

REFERENCES

[1] Singh B, Wongmahesak K, Chandra S. Content marketing with search engine optimization: Fostering successful businesses in online marketing 6.0. Practical Strategies and Case Studies for Online Marketing 6.0, 2025, 1(1): 17-36.

[2] Mager A, Norocel O C, Rogers R. Advancing search engine studies: The evolution of Google critique and intervention. Big Data & Society, 2023, 10(2): 20539517231191528.

[3] Wang Z, Bu D, Wang N, et al. An empirical study on bugs in JavaScript engines. Information and Software Technology, 2023, 155(1): 107105.

[4] Makrydakis N. SEO mix 6 Oâs model and categorization of search engine marketing factors for websites ranking on search engine result pages. Int. J. Res. Mark. Manag. Sales, 2024, 6(1): 18-32.

[5] Ladanavar S M, Kamble R, Goudar R H, et al. Enhancing User Query Comprehension and Contextual Relevance with a Semantic Search Engine using BERT and ElasticSearch. EAI Endorsed Transactions on Internet of Things, 2024, 10(1): 1.

[6] Adnyana I G, Dirgayusari A M, Atmaja K J. Data Visualization for Building a Cyber Attack Monitoring Dashboard Based on Honeypot. Sinkron: jurnal dan penelitian teknik informatika, 2024, 8(4): 2510-2518.

[7] Tkachenko O, Chechet A, Chernykh M, et al. Scalable Front-End Architecture: Building for Growth and Sustainability. Informatica, 2025, 49(1): 1.

[8] Wang J, Huang J X, Tu X, et al. Utilizing bert for information retrieval: Survey, applications, resources, and challenges. ACM Computing Surveys, 2024, 56(7): 1-33.

[9] He X, Liu Q, Jung S. The impact of recommendation system on user satisfaction: A moderated mediation approach. Journal of Theoretical and Applied Electronic Commerce Research, 2024, 19(1): 448-466.

[10] Fan J, Khan J, Singh N P, et al. Fulgor: a fast and compact k-mer index for large-scale matching and color queries. Algorithms for Molecular Biology, 2024, 19(1): 3.

[11] Merlin N R G, Prem. M V. Efficient indexing and retrieval of patient information from the big data using MapReduce framework and optimisation. Journal of Information Science, 2023, 49(2): 500-518.

[12] Soltanmohammadi E, Dilek A, Hikmet N. Tailored Partitioning for Healthcare Big Data: A Novel Technique for Efficient Data Management and Hash Retrieval in RDBMS Relational Architectures. Journal of Data Analysis and Information Processing, 2024, 13(1): 46-65.

[13] Kjorveziroski V, Filiposka S. Webassembly as an enabler for next generation serverless computing. Journal of Grid Computing, 2023, 21(3): 34.

[14] Putra S R, Moraes F, Hauff C. Searchx: Empowering collaborative search research. The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, 1(1): 1265-1268.

[15] Dai S, Li K, Luo Z, et al. AI-based NLP section discusses the application and effect of bag-of-words models and TF-IDF in NLP tasks. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 2024, 5(1): 13-21.

[16] Yang H J, Choi I Y. Enhancing Search Functionality for Website Posts and Product Reviews: Improving BM25 Ranking Algorithm Performance Using the ResNet-Transformer Model. Journal of The Korea Society of Computer and Information, 2024, 29(11): 67-77.

[17] Dejonckheere F. Automated Microservice Identification in Modular Monolith Architectures. 2024, 101(1): 4.

[18] Putra S R, Grashoff K, Moraes F, et al. On the Development of a Collaborative Search System. DESIRES, 2018, 1(1): 76-82.

[19] Zhou Y J, Yao J, Dou Z C, et al. Dynamicretriever: A pre-trained model-based ir system without an explicit index. Machine Intelligence Research, 2023, 20(2): 276-288.

[20] Paulsen D, Govind Y, Doan A H. Sparkly: A simple yet surprisingly strong TF/IDF blocker for entity matching. Proceedings of the VLDB Endowment, 2023, 16(6): 1507-1519.

[21] Gomes L, da Silva Torres R, CÃ´rtes M L. BERT-and TF-IDF-based feature extraction for long-lived bug prediction in FLOSS: a comparative study. Information and Software Technology, 2023, 160(1): 107217.

[22] Danyal M M, Khan S S, Khan M, et al. Sentiment analysis of movie reviews based on NB approaches using TF-IDF and count vectorizer. Social network analysis and mining, 2024, 14(1): 87.

[23] Kungumaraj E, Lathanayagam E, Saikia U, et al. Neutrosophic Topological Vector Spaces and its Properties. International Journal of Neutrosophic Science (IJNS), 2024, 23(2): 63.

[24] Chung Y, Kraska T, Polyzotis N, et al. Automated data slicing for model validation: A big data-ai integration approach. IEEE Transactions on Knowledge and Data Engineering, 2019, 32(12): 2284-2296.

[25] De Falco A, Caruso F, Su X D, et al. A variational algorithm to detect the clonal copy number substructure of tumors from scRNA-seq data. Nature Communications, 2023, 14(1): 1074.

[26] Zhou X, Shen X, Liu Z, et al. A novel shadow calculation approach based on multithreaded parallel computing. Energy and Buildings, 2024, 312(1): 114237.

[27] Xiao F, Wen J, Pedrycz W, et al. Complex evidence theory for multisource data fusion. Chinese Journal of Information Fusion, 2024, 1(2): 134-159.

[28] Manoharan J S. A novel load balancing aware graph theory based node deployment in wireless sensor networks. Wireless Personal Communications, 2023, 128(2): 1171-1192.

[29] Roberts K, Alam T, Bedrick S, et al. Searching for scientific evidence in a pandemic: An overview of TREC-COVID. Journal of Biomedical Informatics, 2021, 121(1): 103865.

[30] Chen J S, Hersh W R. A comparative analysis of system features used in the TREC-COVID information retrieval challenge. Journal of Biomedical Informatics, 2021, 117(1): 103745.

Multi-Criteria Using Dijkstra's Algorithm to Determine Optimal Time Paths in Vehicle Route Optimization

Basorudin[1]*, Handaru Jati[2], Nurkhamid[3], Puput Dani Prasetyo Adi[4]
Doctoral Program of Engineering-Faculty of Engineering, Yogyakarta State University, Sleman Regency, Special Region of Yogyakarta 55281, Indonesia[1]
Department of Electronics and Informatics Engineering Education-Faculty of Engineering, University Negeri Yogyakarta, Sleman Regency, Special Region of Yogyakarta 55281, Indonesia[2, 3]
National Research and Innovation Agency (BRIN), Bandung, Indonesia[4]

***Abstract*—This research is the development of a Road Traffic Network using one of the methods in Mathematics, namely Dijkstra's Algorithm, Weighted Sum Method (WSM), and Weighted Product Method (WPM). Meanwhile, the parameters used are route, volume, capacity, DOS, distance, and travel time. The objective of this research is to find the fastest route alternative from one place to another. The recommended results using Dijkstra's algorithm combine values of distance, travel time, and congestion degree with the weighted-sum and weighted-product methods, each calculated accordingly. The shortest route is 1→2→4→7→10→14, the route with the least congestion and shortest travel time is 1→2→5→6→9→13→14. This research combines these three parameters to obtain a balanced route between congestion level, short travel distance, and short travel time for the driver. By combining these parameters, the best route from this study is route 1: 1→2→5→8→12→11→14 with a total distance of 28.77 km, a saturation degree value of 5.421, and a travel time of 28 minutes. Thus, the research results indicate that the best route will have a combination of multiple criteria, such as short distance, short travel time, and less congestion simultaneously. The Weighted-Sum Method (WSM) and Weighted-Product Method (WPM) can produce different outputs, with WPM being superior to WSM in terms of computational steps.**

***Keywords*—Dijkstra's algorithm; multi-criteria; road traffic network; Weighted Sum Method (WSM); Weighted Product Method (WPM); mathematics**

I. Introduction

Currently, in solving shortest paths, several complicated mathematical methods are used, and certainly need to be efficient, one of which is wCDN.The proposed algorithms are MRD labeling and MRCCSPP. The route is evaluated at confidence levels of 0.45 and 0.85. Three criteria: distance, cost, and time are used. Paths are calculated for all criteria. Simulation shows the efficiency of the MRD algorithm for larger examples [1]. The research conducted by Syarif Surorejo et al. (2024) discusses the comparison between Dijkstra's algorithm and the genetic algorithm to determine the distance between points on two alternative routes to Guci. The results of the study indicate that Dijkstra's algorithm is superior in this case, but in general, the genetic algorithm has potential for more complex optimization problems [2]. This research will focus on the development of the study conducted by Noraini Abdullah & Ting Kien Hua (2018) on transportation issues. To reduce transportation costs related to the shortest path problem by applying Dijkstra's algorithm, extensive research has been conducted, including a study by M. Afdhaluzzikri et al. (2024). [3]. Besides transportation, even the search for the shortest route in the case of election logistics distribution has also been carried out. Based on the results of this study, the shortest route for distributing election logistics from the Kesugihan District Office to each Village Head Office in Kesugihan District was found [4].

Meanwhile, the research by Yann Disser et al., titled "Multi-criteria Shortest Paths in Time-Dependent Train Networks", used a synthetic dataset with 1,000 random test cases. Realistic dataset with 1,000 one-hour interval tests. Real dataset from Deutsche Bahn AG with 14,000 connection queries. Train schedule data from Germany, 2007 (56,994 trains). With multi-criteria generalization of Dijkstra's algorithm, a speedup factor of 20 for label creation was achieved. A speedup factor of 138 for label insertion was achieved. This new approach is competitive with the time-extended method. Performance improves with the addition of more criteria [5], [49].

Dijkstra's algorithm is one of the algorithms used to solve the shortest path and does not have a negative cost [6]. The Dijkstra algorithm is effectively used to find the shortest and most cost-effective routes for intercity flights, demonstrating a high success rate in identifying the best paths for intercity transportation in a shorter time and at a lower cost. Through simulations, the algorithm demonstrated a 95% success rate in quickly determining the most suitable flight routes, regardless of the number of cities involved in the network, highlighting its efficiency and reliability in optimizing flight routes [7]. Meanwhile, the research related to Dijkstra's Algorithm for Determining the Shortest Route from Surabaya City Center to Historical Sites aims to find the shortest route from the city center of Surabaya to historical sites using Dijkstra's algorithm, which helps reduce travel time [12] and is efficient for visitors. The Dijkstra algorithm was chosen for its ability to determine the shortest path in a weighted graph with positive weights, starting from one point to all desired points. This study found five shortest routes starting from Gubeng Station,

*Corresponding author.

demonstrating the practical application of Dijkstra's algorithm for tourists and locals visiting historical sites. Previous research has also applied Dijkstra's algorithm in various contexts, such as water meter reading routes and school bus routes, demonstrating its versatility in optimizing travel and costs [8].

The research conducted by Israa Ezzat Salem and Maad M. Mijwil shows that Dijkstra's Algorithm is effectively used to find the shortest and most cost-effective routes for intercity flights, demonstrating a high success rate in identifying the best routes for intercity transportation in a shorter time and at a lower cost [46]. Through simulations, the algorithm demonstrated a 95% success rate in quickly determining the most suitable flight routes, regardless of the number of cities involved in the network, highlighting its efficiency and reliability in route optimization for flights [9]. The research conducted by Asrahmaulyana et al. (2020), titled "Economic Losses Due to Traffic Congestion on the Muslim Community Residing in Makassar City", concludes that the losses due to congestion include expenditures for fuel purchases under normal conditions, amounting to Rp 19,285.38 for car users and Rp 7,217.86 for motorcycle users. However, if they get stuck in traffic, the costs increase to Rp 29,428.46 per car and Rp 13,175.71 per motorcycle [10],[11].

Research related to Based Shipping Routes Using the Dijkstra Algorithm has introduced a method to identify and characterize shipping routes using data from the Automatic Identification System (AIS), which ships use to broadcast their location and movements to coastal traffic services. This method involves creating a graph where the sea area is divided into grid cells, and the movement of ships between these cells is tracked to find the most frequently used routes, which are assumed to be the safest. By applying Dijkstra's algorithm to this graph, the research successfully identified potential safe routes between two locations, which are the most frequently used paths by ships. This approach helps in determining the safest and most efficient route for maritime navigation [6].

Meanwhile, the article titled "Least Congested And Shortest Distance Path In Kota Kinabalu Traffic Network, Sabah, Malaysia" has focused on traffic estimation using GPS devices installed in vehicles to predict current traffic conditions and suggest the least congested alternative routes (Ting Kien Hua & Noraini Abdullah, 2017). The research conducted by Ting Kien Hua and Noraini Abdullah titled "Weighted Sum-Dijkstra's Algorithm in Best Path Identification based on Multiple Criteria" discusses how GPS devices in vehicles have been used for traffic estimation to predict current traffic conditions and suggest the least congested alternative routes. In this study, the Dijkstra Algorithm has been applied in maritime networks to find paths based on parameters beyond distance, enhancing the route selection process [7]. The solution proposed in this article suggests using Dijkstra's algorithm with multi-criteria to optimize vehicle routes [48]. The Weighted Product Model (WPM). The WPM is very similar to the WSM. The main difference is that instead of addition in the model, there is multiplication. Each alternative is compared to the others by multiplying several ratios, one for each criterion.[13]. Some of the objectives of this research include considering the impact of uncertain traffic conditions on driving speed. By finding the best route, travel time can be shortened [14]. Traffic congestion can cause various negative impacts [15]. The shortest path in computer networks is a routing category that is generally used in computer networks today.[16].

The article titled "Optimal Routes in Urban Networks Using Travel Time Prediction" proposes an algorithm that determines the optimal route between the starting point and the endpoint on the road network. The goal is to minimize travel time while considering constraints such as maximum travel time and risk level [17]. Whereas the article titled "A Path Planning and Navigation Control System Design for Driverless Electric Bus" defines the concept of "path" referring to the process of planning and creating a route for a driverless bus, with the main components involved in route planning are Global Route Planning, local route planning, track creation, special routes and maneuvers, following the route and error correction [18].

The problem in Vehicle Route Optimization is very complex. The research conducted by Tegar Arifin Prasetyo (2021) explains that the problem of "Routing" or vehicle routing is approached using optimization techniques. This research successfully demonstrates how optimization methods can be applied in the real world to address logistics distribution problems by considering various constraints such as vehicle capacity, operational costs, and delivery time [19]. In urban areas, where infrastructure and population density enable extensive data collection and real-time service optimization, routing solutions can be implemented effectively [20].

This research is closely related to expert systems and Artificial Intelligence (AI) because the scientific foundation of Dijkstra's algorithm is Expert Systems and Artificial Intelligence (AI). In the article written by Jonhariono Sihotang (2020). Kecerdasan Buatan (AI) is described as a branch of computer science that focuses on automating intelligent behavior. AI enables computers to perform tasks that are usually performed by humans [21]. The use of artificial intelligence (AI) technology in organizations is becoming increasingly urgent. AI can automate processes, reducing reliance on manual intervention and increasing the speed of data processing and decision-making [22].

Whereas an expert system refers to a computer-based system designed to mimic the ability to think and make decisions of an expert or specialist in the diagnosis of sexual disorders (paraphilia) [23]. An expert system is a computer application that uses knowledge and techniques that are usually possessed by an expert to solve problems in a specific field [24]. Expert systems not only help in solving optimization problems but can also address other cases, such as in the article by Taozara Laia et al. (2022), where expert systems are used to diagnose eye diseases caused by excessive computer use, known as Computer Vision Syndrome (CVS) [25].

This research analyzes Dijkstra's algorithm for solving shortest paths by comparing other similar research and finding a point of novelty, by completing the entire research design using algorithms that are more complete than before, some of which are the Weighted Sum Method (WSM) and the Weighted Product Method. Method development is necessary to improve effectiveness, in addition to novelty, and also a higher level of precision [52],[53].

II. THEORY

A. Dijkstra's Algorithm

Research conducted by Angshuman Guin, et al. about travel time measurement uses three types of travel time measurements: (a) Instantaneous Travel Time (ITT): calculated based on the average instantaneous speed on each road segment, (b) Reactive Travel Time (RTT): based on the travel time of vehicles that have just exited the corridor, and (c) Actual Travel Time (ATT): measured through Bluetooth data, which provides actual travel time in the field. The ITT, RTT, and ATT data were analyzed for each 5-minute interval during the evening congestion period on the I-285 corridor in Georgia. The comparison between ITT, RTT, and ATT shows that simulation-based predictions are more consistent with actual travel time (ATT), especially during congestion [34].

In graph theory, Dijkstra's algorithm is a graph traversal algorithm introduced by Dutch computer scientist Edsger W. Dijkstra in 1959. The Dijkstra algorithm is also known as the single-source shortest path algorithm, similar to Prim's algorithm. Using Dijkstra's algorithm to find the shortest path in a directed graph with non-negative weights is one of the fundamental issues in algorithmic problems. Peter W. Eklund introduced a modified Dijkstra algorithm that includes static and dynamic components, which are simultaneously used to route emergency vehicles in an earthquake simulation in Okayama, Japan. In the article by Peyer, S., Rautenbach, D., and Vygen, J, a new algorithm called Generalized Dijkstra is introduced, which is a fast technique for the Dijkstra algorithm [35]. Penelitian yang dilakukan oleh Ramiz ASSAF, dkk. (2017), tentang Vehicle-Routing Optimization For Municipal Solid Waste Collection Using Genetic Algorithm: The Case Of Southern Nablus City.

B. Degree of Saturation (DoS)

DOS or traffic density is an indicator of traffic flow with a range between 0 and 1. The range of DOS values is between 0 and 1. A saturation degree of less than 0.85 essentially indicates that the capacity is still sufficient for the traffic volume demand. However, when the density approaches 1.0, traffic flow may become unstable, with queues and delay conditions occurring. When demand exceeds capacity (density greater than 1.0), traffic will experience congestion, long queues, and excessive delays. Here is the DOS as Eq. (1):

$$Degree\ of\ Saturation = \frac{Volume}{Capacity} \quad (1)$$

C. Weighted Sum Method (WSM)

The simple WSM method is the summation of these variables. This implies that there is a level of trade-off remaining between each pair of criteria. The most commonly used normalization method is dividing the values by the total of all criterion values to obtain a proportion of the whole. After normalization, assign weights to the criteria and sum the weighted values to produce the appropriate scores and rankings. The WSM equation is shown in Eq. (2):

$$Weighted\ Sum\ Method = W_1X_1 + W_2X_2 + \cdots + W_nX_n \quad (2)$$

D. Weighted Product Method (WPM)

Yoon in Kusumadewi et al. (2006) explains the WPM method using multiplication to link attribute ratings, where the rating of each attribute must first be raised to the power of the corresponding attribute weight. This process is similar to the normalization process. The WPM method is a straightforward and flexible approach because it can directly ignore normalization. The weight seems to be an exponent of the criteria. WPM is shown as in Eq. (3):

$$Weighted\ Product\ Method = {W_1}^{w1}{W_2}^{w2} \ldots\ldots {W_n}^{wn} \quad (3)$$

E. Time Travel

The article written by (Hyungjoo KIM, et al. 2013) discusses the concept of travel time. This article focuses on the travel time estimation methods used to monitor vehicle speeds on highways, particularly in South Korea. Sample-Travel Speed is measured by re-identifying vehicles using transponder electronic toll collection (ETC), which allows for accurate travel time measurements over a certain distance [39].

Meanwhile, another study on travel time estimation using the Instantaneous Model estimation method with the local speed survey method (Spot speed) has been conducted, and the results of this study estimate the travel time for vehicles crossing the studied road segment during the first 5-minute interval from 12:00 to 12:05 WIB from upstream link 1 to downstream link 3 as 2.346 + 2.498 + 2.521 = 7.364 minutes [40].

F. Path

The article, titled "Optimal Paths in Urban Networks Using Travel Time Predictions", proposes an algorithm that determines the optimal path between the starting point and the end point in a road network. The goal is to minimize travel time by considering constraints such as maximum travel time and risk level. This study shows that the proposed algorithm efficiently finds the optimal path for urban networks, taking into account travel time and risk constraints [41].

Meanwhile, the article entitled "A Path Planning and Navigation Control System Design for Driverless Electric Bus" defines the concept of "path" as referring to the process of planning and creating a path for a driverless bus, with the main components involved in path planning being global path planning, local path planning [36], path creation, special paths and maneuvers, path following and error correction. This entire process ensures that driverless buses follow a safe, smooth, and efficient path while adapting to dynamic road conditions. [42].

G. Distance

In the article "AIS-Based Shipping Routes Using Dijkstra's Algorithm", the term "distance" refers to the weights of the sides in a graph that represent the movement of ships between cells in a grid. This weight is inversely proportional to the frequency of ship movement between cells, which means the more often the ship moves between two grid cells, the lower the weight (or "distance") between those cells in the graph [43].

H. Route

When traveling, the question often arises as to which route is the fastest. In answering this question, the criterion that is

often used is distance because the shorter the distance, the faster the travel time. In research conducted by Tetra Widianto, et al. The best route is one that considers the shortest distance and the availability of public facilities. The needs in question are user preferences for available public facilities. Based on the two scenarios that have been run, the resulting routes are the best in each scenario, calculated using Dijkstra's algorithm with a weighted sum input [44].

I. Calculation of Vehicle Volume

Research conducted by Hj Indra Harun discusses the effect of traffic volume on the noise level at the Bastiong highway intersection. The calculation of vehicle volume based on time was carried out at 3 different points for 10 hours, as seen in Table I, Table II, and Table III. Based on the LHR calculation using formula 2, it shows that the highest vehicle volume occurs at point A during the day. Based on the vehicle volume table at point A, the increase in the number of vehicles occurs at 10:00 WIT, this is influenced by police activities that are conducting ticketing operations at 07:00 - 10:00 WIT around the research location [37], against drivers who violate traffic rules, especially two-wheeled vehicles, so that many two-wheeled drivers do not pass through this road segment in the morning. 00 WIT in the vicinity of the research location, against drivers who violate traffic rules, especially two-wheeled vehicles, so that many two-wheeled drivers do not pass through this road section in the morning [45].

TABLE I. COMPARISON BETWEEN LV, HV, AND MC

Time	Vehicle Type			
	LV (Samples/hours)	*HV (Samples/hours)*	*MC (Samples/hours)*	*(Samples/hours)*
07.00-08.00	447,00	34,80	452,25	934,05
08.00-09.00	435,00	66,00	421,75	922,75
09.00-10.00	345,00	18,00	427,50	790,50
10.00-11.00	333,00	48,00	556,25	937,25
11.00-12.00	529,00	116,40	631,50	1276,90
12.00-13.00	500,00	66,00	587,50	1153,50
13.00-14.00	394,00	84,00	524,00	1002,00
14.00-15.00	529,00	122,40	555,25	1206,65
15.00-16.00	459,00	121,20	610,50	1190,70
16.00-17.00	593,00	135,60	758,50	1487,10
Total	4564,00	812,40	5525,00	10901,40

Data source: Disser, Y., Müller-Hannemann, M., & Schnee, M. (2008).

III. METHODS

A. Solving the Routing Problem

The Relevant articles are identified using the keywords "Dijkstra", "Shortest Path", "Multi Criteria", "Weighted Sum Method", "Weighted Product Method", "Distance", "Optimal Path", "Routing Problem", "Time Travel", Degree of Saturation (DOS), or Density. After being identified, the papers were classified using the framework developed by [26]. Traffic congestion is a common problem that requires traffic compliance [27]. The research conducted by Nurwan et al. (2021) discusses the optimal bus schedule starting from 6:30 AM to 5:00 PM. Each bus has 4 (four) departure sessions and 4 (four) return sessions, with a travel time of 60 minutes for each session. The research results show that the shortest route was obtained $V1 - V2 - V5 - V8 - V9 - V10 - V13 - V16$ [28].

Moreover, Traffic congestion in the city of Pekanbaru occurs due to the high number of vehicles on the road and the slow improvement of road facilities; for example, the road capacity remains the same while the number of vehicles continues to increase. Traffic congestion can be categorized into two types: recurring congestion and non-recurring congestion. Recurring congestion is a type of congestion that includes excess demand for travel on four-wheeled and two-wheeled vehicles, as well as a shortage of intersection supply on every road. Recurring traffic jams usually occur during peak hours or working hours, especially from 07:00 AM to 09:00 AM, and from 04:30 PM to 07:00 PM. This is caused during peak hours when people tend to go to work, children go to school, and parents drop off and pick up their children from school, or workers return from their workplaces to go home. Another type of traffic congestion is non-recurring congestion caused by specific cases such as accidents or road closures. The literature review in this article will detail the modeling procedures and their outputs for comparing various traffic parameter input scenarios [50], [51].

B. Research Design

The research conducted refers to the previous study by Dr. Noraini Abdullah & Ting Kien Hua (2018) [29]. With different data and location, previous research created the following research design, as shown in Fig. 1.

Flowchart or research design in Fig. 1. Conducted by Dr. Noraini Abdullah & Ting Kien Hua (2018), it can be explained as follows: traffic data was collected and prepared in Kota Kinabalu during peak hours from 5:00 PM to 7:00 PM. From the data, there are five different input parameters, namely distance, DOS or density, travel time, WSM combined parameter method, and WPM. With all the collected data, five different types of directed network graphs were formed. Then, the traffic network was solved using the Dijkstra algorithm and

both methods. The results of the five different parameters were then compared. Finally, the combined parameter with WSM or the equilibrium between all criteria will be shown in the results section.

The development of the research that will be conducted from the previous study by Dr. Noraini Abdullah & Ting Kien Hua (2018) [30],[31],[32] includes innovations by adding variables related to Time Travel, which previously only considered travel time. It will later be developed to include the fastest time, longest time, and average time. Additionally, there will be the addition of variables related to Elevation or Altitude by calculating the road slope data obtained from Google Earth Pro or the Google API. Furthermore, there will be the addition of traffic lights at intersections, which will be counted in terms of the number of red lights and the duration of the red lights (in seconds). The research conducted by Hyungjoo KIM, et al., about Assessment of Travel Time Estimates based on Different Vehicle Speed Data: Spot Speed vs. Sampled Journey Speed in South Korean expressways, compares two methods of measuring vehicle speed on South Korean highways, namely instantaneous speed measured with loop detectors and actual journey speed measured with electronic toll collection (ETC) transponders. The research was conducted on the Gyeongbu expressway section from Anseong to Giheung-Dongtan, covering a distance of 26.4 km. Speed data were collected through 23 loop detector stations that measure average time speed (instant time at a fixed location) and 8 ETC reader points that measure average distance speed (calculated from travel time between readers). The results of this study indicate that loop detectors are less sensitive to congestion compared to the ETC method [33].

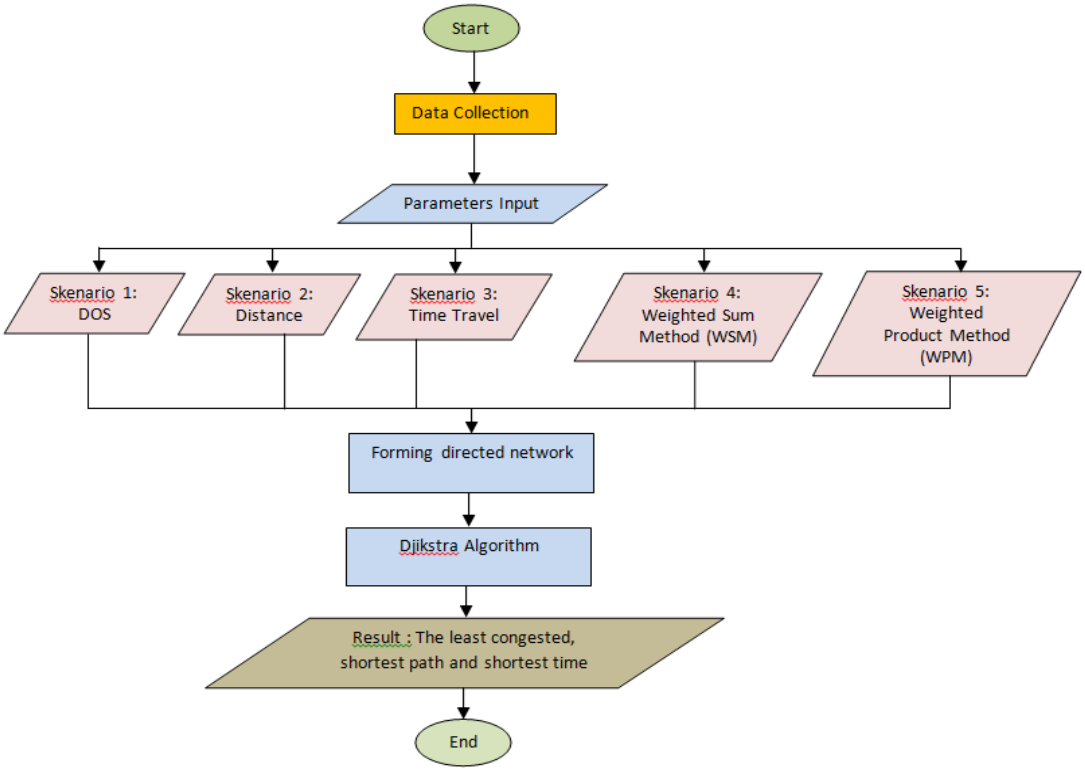


Fig. 1. Research design.

IV. Result and Analysis

A. Research Location

In previous research conducted by [47], the scope selected in the study was the network from Inanam, a suburban city (source node), to Centre Point, which forms the Central Business District (CBD) in the city of Kota Kinabalu (receiver node). Meanwhile, the research conducted in this study is in the city of Pekanbaru, Riau Province, Indonesia. All routes between the source node and the receiving node of the coverage are defined. In Fig. 2, the yellow dot is part of the main intersection selected in the city of Pekanbaru, Riau Province, Indonesia, which is defined as the network graph node.

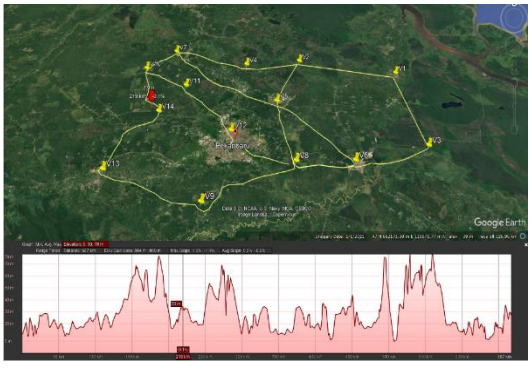


Fig. 2. Traffic network in Pekanbaru City, Riau Province, Indonesia.

B. Data Collection

Data, as shown in Table II, was collected on the road and through Google Maps. Lane distances were collected from Google Maps and relevant agencies, for example, on the official website https://www.pekanbaru.go.id/ [38]. Volume, saturation distance, cycle time, green time, and intersection geometry were collected to calculate the density level (volume/capacity ratio). The time required was collected through real-time traffic flow on Google Maps and Google Earth during the afternoon rush hour.

In Table II, it can be explained that V is a node that is a path from node to node. V1 to V14 are nodes or intersections in the traffic network in Pekanbaru, Riau Province, which were taken with the tools of Google Earth. 'Volume' is the number of vehicles occupying a particular lane at a particular time. 'Capacity' is the maximum volume of vehicles that can pass through a particular intersection or lane. Furthermore, the Degree of Saturation (DOS) is the volume/capacity ratio, which is an indicator of current traffic conditions. 'Distance' is the length of the path from node to node. The 'Weighted Sum Value' is the combination of degree of saturation, travel time, and distance using the Weighted Sum Methods (WSM). In addition, the 'Value of Weighted Product Methods (WPM) is the combined value of several criteria using the Weighted Product Methods (WPM).

C. Dijkstra's Algorithm Output

The network graph in Fig. 3 has a degree of saturation (DOS) parameter. Degree of saturation is the ratio of volume to capacity. The yellow-highlighted path in Fig. 3 below is the path with the smallest degree of saturation (DOS = 1.003+1.006+0.126+0.087+0.409), which indicates the least congested path, namely $V1 \rightarrow V2 \rightarrow V4 \rightarrow V7 \rightarrow V10 \rightarrow V14$ with a total DOS value of 2.631.

TABLE II. COMPARISON OF THE DATA COLLECTED

Path	*From*	*To*	*Volume*	*Capacity*	*DOS (Value)*	*Distance (Km)*	*Time Travel (minute)*	*Weighted Sum Method (WSM) (Value)*	*Weighted-Product Method (WPM)(Value)*
1	V1	V2	191482	191000	1,003	4,25	4	0,021338	2,573442702
2	V1	V3	3338	3238	1,031	8,54	8	0,032522	4,129705815
3	V2	V4	87481	87000	1,006	8,32	7	0,030575	3,883338819
4	V2	V5	929431	929231	1,000	5,16	5	0,023837	2,955092343
5	V5	V2	794	1798	0,442	5,16	6	0,022280	2,711023624
6	V3	V6	44432	43432	1,023	5,39	5	0,024357	3,020982905
7	V4	V7	145	1152	0,126	18,57	16	0,046586	3,344124748
8	V7	V5	27750	26750	1,037	19,07	17	0,058156	6,954185511
9	V5	V7	537034	527030	1,019	19,07	10	0,040923	5,124951434
10	V8	V5	360	1302	0,276	5,60	5	0,016794	1,978259148
11	V5	V8	39596	37466	1,057	5,60	4	0,020076	2,265351286
12	V6	V5	1081	1475	0,733	22,94	17	0,059691	6,587065817
13	V5	V6	26251	27151	0,967	22,94	2	0,016089	1,706505584
14	V6	V9	365454	36800	9,931	1,77	2	0,109002	3,275890584
15	V8	V12	302	1162	0,260	1,85	2	0,007805	0,987038697
16	V12	V8	27009	25004	1,080	1,85	10	0,039843	5,032614609
17	V7	V10	145	1660	0,087	9,60	8	0,023933	1,88599431
18	V10	V11	19097	10287	1,856	9,10	8	0,041852	5,131794566
19	V11	V10	340967	310025	1,100	9,10	6	0,029381	3,702238625
20	V11	V12	146	2308	0,063	3,60	3	0,009294	0,880738407
21	V12	V11	26271	25094	1,047	3,60	2	0,015621	1,464454856
22	V12	V9	161	200	0,805	6,76	6	0,025158	3,196180338
23	V9	V12	15194	16924	0,898	6,76	17	0,055632	6,524738915
24	V9	V13	300584	177785	1,691	12,45	12	0,049856	6,321310361
25	V13	V14	206	250	0,824	3,55	2	0,015789	1,801891017
26	V11	V14	21365	20265	1,054	8,30	2	0,024000	2,596299516
27	V10	V14	126	308	0,409	53,00	50	0,139605	10,27279372
28	**Amount**		**3006202**	**2506097**	**31,8246**	**273,06**	**236**	**1,000000**	**100,3080083**

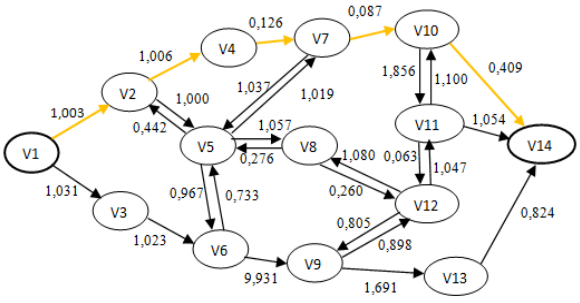


Fig. 3. Output of Dijkstra's algorithm (degree of saturation parameter network graph).

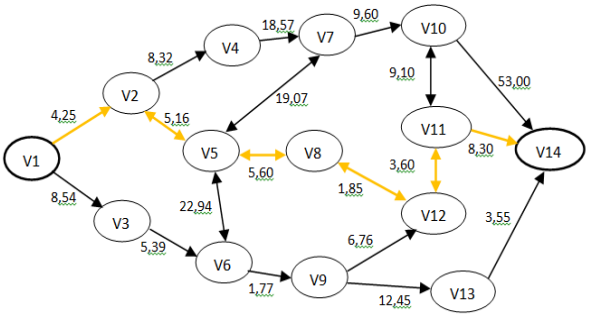


Fig. 4. Output of Dijkstra's algorithm (distance parameter network graph).

The Network Graph in Fig. 4 with distance parameters is formed using the data collected in Table II. The source node is node 1, and the receiving node is node 14. From Fig. 4, the path marked in yellow is represented by 1→2→5→8→12→11→14, which is the shortest path with a total distance (4.25+5.16+5.60+1.85+3.60+8.30) equal to 28.77 kilometers.

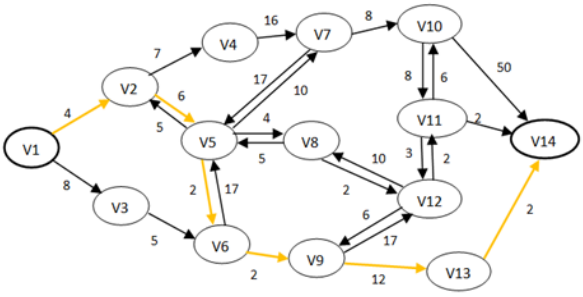


Fig. 5. Output of Dijkstra's algorithm (time travel parameter network graph).

In Fig. 5, the time travel parameter network graph is shown together as an alternative path. The alternative path with the shortest travel time is path 1→2→5→6→9→13→14. The total travel time is 28 minutes, traveling from 1 to 14, respectively.

In Fig. 6 and Fig. 7, the Dijkstra algorithm uses a combination of distance, travel time, and density values with the weighted-sum and weighted-product methods, respectively, executed. From the output in Fig. 6, the shortest distance and travel time is 1→2→4→7→10→14, while the output in Fig. 7 is the path with less density, shorter distance, and shorter travel time, namely the path 1→2→5→6→9→13→14. The total distance of this alternative route is 28.77 kilometers with a total saturation degree of 5.421 and a travel time of 28 minutes.

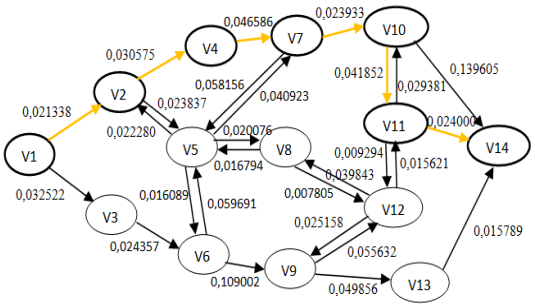


Fig. 6. Output of Dijkstra's algorithm (weighted-sum method).

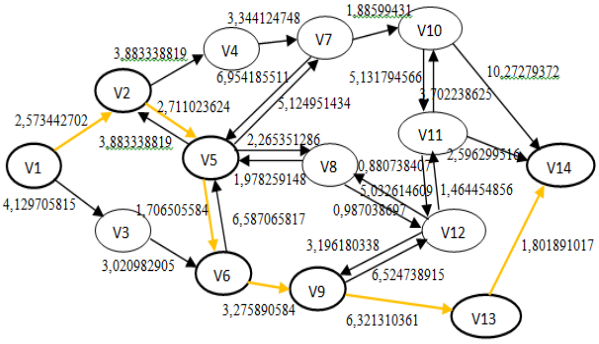


Fig. 7. Output of Dijkstra's algorithm (weighted-product method).

TABLE III. COMPARISON OF DISTANCE, DENSITY, AND TRAVEL TIME

Parameter	Specifics Parameter			
	Path	Distance (km)	DOS (Degree of Saturation) (Density)	*Time Travel (Travel Time/Minute)*
Distance	Path 1: 1->2->5->8->12->11->14	28,77	5,421	28
DOS	Path 2: V1->V2->V4->V7->V10->V14	93,74	2,631	85
Time Travel	Path 3: 1->2->4->7->10->14	50,12	15,416	28
Weighted SUM Method	1->2->4->7->10->14	58,14	2,631	85
Weighted Product Method	1->2->5->6->9->13->14	50,12	5,421	28

Data source: Personal Research Data

As shown in Table III, there are three different paths selected depending on the selected parameters. With the distance parameter network graph, Path 1 is the path selected with the shortest distance. Path 2 is the path selected with Degree Of Saturation (DOS), while Route 3 is the route with the shortest travel time. With the combined parameters of DOS, travel time, and distance, Route 1 was chosen because it has achieved a balance between distance, travel time, and DOS. In other words, the driver can decide on the optimal right lane that is not too congested, has a shorter travel time, and a shorter distance. The difference between the paths in percentage is then calculated to find significant differences between them.

V. Conclusions

The shortest distance route may not be the best; two different parameters result in two different routes, but the best route is chosen based on distance and the lowest saturation level criterion. However, in this study, travel time is an additional criterion. Routes with three different parameters have been compared and result in three different routes. From these results, the shortest route is 1→2→4→7→10→14, the least congested route, and the shortest travel time is 1→2→5→6→9→13→14. This study combines these three parameters to get a balanced route between congestion level, short travel distance, and short travel time for drivers. By combining these parameters, the best path from this study is path 1: 1→2→5→8→12→11→14 with a total distance of 28.77 km, a saturation degree value of 5.421, and a travel time of 28 minutes. Thus, the results show that the best path will have a multi-criteria combination of short distance, short travel time, and fewer simultaneous traffic jams. The Weighted-Sum Method (WSM) and the Weighted-Product Method (WPM) can produce different outputs, with WPM being superior to WSM in terms of computational steps. The Weighted Product Method can also be user-friendly, and avoids normalization issues that can cause the reversal of rankings usually faced by the Weighted-Sum Method.

VI. Future Research

The calculation results on Time Travel on the graph become three (3) graphs, namely the fastest time, the longest time, and the average time in minutes, while altitude affects speed. The more sloping the road conditions, the longer it will take, and it is not recommended, because it will affect not only costs but can also affect time and reduce the bad energy caused by vehicle exhaust fumes. Then the number of red lights will also affect the time of arrival at the destination. This is calculated based on the number of red lights and the length of time in seconds on the red light indicator on the road.

Acknowledgment

Thanks to all internal and external teams who have helped to complete this research. Hopefully, this research will continue to develop in the future so that it can contribute to the development of smart cars and the like.

References

[1] Abdullah, N., & Hua, T. K. (2018). Weighted Methods of Multi-Criteria Via Dijkstra's Algorithm in Network Graph For Less Congestion, Shorter Distance, and Time Travel in Road Traffic Network. NETWORK GRAPH FOR LESS CONGESTION, SHORTER DISTANCE, AND TIME. July. https://doi.org/10.5281/zenodo.1305010.

[2] Surorejo, S., Muhammad Raikhan Al. F, Wresti A. & Gunawan (2024). Comparison of Dijkstra and genetic algorithms for shortest path guci, Journal Mandiri IT ISSN 2301-8984 (Print), 2809-1884 (Online) Vol. 13 No. 1, July (2024), pp. 56-62. Journal homepage: www.ejournal.isha.or.id/index.php/Mandiri.

[3] M. Afdhaluzzikri, L. D. Santriawan, Ermawati S. M., Setyo N., & Mamika U. R. (2024). Application of Dijkstra Algorithm in Determining Transportation Costs for Tourist Attractions in Lombok Island Based on the Shortest Path. Jurnal Pariwisata Nusantara (Juwita), Volume 3, No. 2, August 2024. Homepage: https://journal.uinmataram.ac.id/index.php/juwita.

[4] Isnaeni, N., Mizan A., & Widayati, R. (2024). Implementation Of The Djikstra Algorithm To Determine Shortest Route For Logistics Distribution Of The 2024 Election In Kesugihan District. Journal of Mathematics Education and Science. VOL. 7 NO. 2 (2024) : 101-107. E-ISSN: 2621-1211.DOI: https://doi.org/10.32665/james.v7i2.1897.

[5] Bunaen Chatrin, Pratiwi Hanna, & Riti Yosefina. (2022). Penerapan Algoritma Dijkstra Untuk Menentukan RuteTerpendek Dari Pusat Kota Surabaya Ke Tempat Bersejarah. Jurnal Teknologi Dan Sistem Informasi Bisnis, 4(1), 213–223.

[6] Sari. P.I., M. Fahri, F., M. ILham M., & Isnaini, F., Q. (2021). Implementation of Dijkstra's Algorithm to Determine the Shortest Route in a City. Journal of Computer Science, Information Technologi and Telecommunication Engineering. Vol.2, No. 1. https://jurnal.umsu.ac.id/index.php/jcositte/article/view/6503.

[7] Chilukuri, B. R., Laval, J. a., Guin, A., Laval, J. a., Chilukuri, B. R., Technology, G. I. of, Transportation, G. D. of, Administration, F. H., (GTI-UTC), G. T. I. U. T. C., & Administration, R. and I. T. (2013). Freeway Travel-time Estimation and Forecasting. Transportation Research Record, 57p. http://www.utc.gatech.edu/sites/default/files/projects/reports/guin_laval__freeway_travel_time_estimation_0.pdf%0Ahttp://g92018.eos-intl.net/eLibSQL14_G92018_Documents/10-22.pdf%0Ahttps://trid.trb.org/view/1246696.

[8] Cutolo, A., De Nicola, C., Manzo, R., & Rarità, L. (2012). Optimal paths on urban networks using travelling times prevision. Modelling and Simulation in Engineering, 2012. https://doi.org/10.1155/2012/564168.

[9] Disser, Y., Müller-Hannemann, M., & Schnee, M. (2008). Multi-criteria shortest paths in time-dependent train networks. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5038 LNCS, 347–361. https://doi.org/10.1007/978-3-540-68552-4_26.

[10] Asrahmaulyana, Qarina, Lindry, E. E., (2020). Economic Losses Due to Traffic Congestion Traffic Against the Muslim Community That Residing in the City of Makassar. Journal Iqtisaduna. Volume 6 Nomor 2 Ed. Desember 2020 : page 157-166 p-ISSN: 2460-805X e-ISSN : 2550-0295 DOI : 10.24252/iqtisaduna.v6i2.19016.

[11] Hj, I. H., & Saputra, M. T. Y. (2021). Pengaruh Volume Lalu Lintas Terhadap Tingkat Kebisingan Disimpang Tiga Jalan Raya Bastiong. Clapeyron : Jurnal Ilmiah Teknik Sipil, L(2), 59–65. https://doi.org/10.33387/clapeyron.v2i2.3904.

[12] Kim, H., Kim, S., Hyoung, S., & Jang, K. (2013). Assessment of Travel Time Estimates based on Different Vehicle Speed Data : Spot Speed vs Sampled Journey Speed in South Korean expressways. Proceedings of the Eastern Asia Society for Transportation Studies, 9, 15. http://easts.info/online/proceedings/vol9/PDF/P334.pdf.

[13] Chourabi. Z., Amel B. D., & Khedher F. (2018). Multi-criteria decision making in workforce choice using AHP, WSM, and WPM. The Journal of The Textile Institute. D OI: 10.1080/00405000.2018.1541434.

[14] Zhao, J., Hu, H., Han, Y., & Cai, Y. (2023). A review of unmanned vehicle distribution optimization models and algorithms. Journal of Traffic and Transportation Engineering (English Edition), 10(4), 548–559. https://doi.org/10.1016/j.jtte.2023.07.002.

[15] Sulistyono. (2022). Kerugian Ekonomi Dan Lingkungan Sebagai Dampak Kemacetan Transportasi Kendaraan Bermotor Pengguna Bbm Fosil. Majalah Ilmiah Swara Patra, 12(2), 12–21. https://doi.org/10.37525/sp/2022-2/274.

[16] Behera, S. P., Bhattacharjee, S., & Mishra, D. (2020). Sustainable Humanosphere Application of Genetic Algorithm for Shortest Path Routing Problem. February.

[17] Cutolo, A., De Nicola, C., Manzo, R., & Rarità, L. (2012). Optimal paths on urban networks using travelling times prevision. Modelling and Simulation in Engineering,2012. https://doi.org/10.1155/2012/564168.

[18] Yu, L., Kong, D., Shao, X., & Yan, X. (2018). A path planning and navigation control system design for driverless electric bus. IEEE Access, 6(c), 53960–53975.https://doi.org/10.1109/ACCESS.2018.2868339.

[19] Prasetyo, T. A. (2021). Particle Swarm Optimization and Genetic Algorithm for Big Vehicle Problem: Case Study in National Pure Milk Company. International Journal of Computing Science and Applied Mathematics, 7(1), 28. https://doi.org/10.12962/j24775401.v7i1.8210.

[20] Porru, S., Misso, F. E., Pani, F. E., & Repetto, C. (2020). Smart mobility and public transport: Opportunities and challenges in rural and urban

areas. Journal of Traffic and Transportation Engineering (English Edition), 7(1), 88–97. https://doi.org/10.1016/j.jtte.2019.10.002.

[21] Sihotang, J. (2020). Analysis Of Shortest Path Determination By Utilizing Breadth First Search Algorithm. Jurnal Info Sains : Informatika Dan Sains, 10(2), 1–5. https://doi.org/10.54209/infosains.v10i2.30.

[22] Silalahi, M. J. A. V.,* , & Wardani. P.F.,. (2024). Application of Artificial Intelligence Technology to Improve Responsiveness and Speed of Operations in Organizations. Prosiding Seminar Nasional Ilmu Manajemen, Ekonomi, Keuangan dan Bisnis Vol. 3, No. 2 (November) 2024: 145-164. DOI: https://doi.org/10.55927/snimekb.v3i2.12645.

[23] Gulo, N. R. (2022). Expert System for Diagnosis of Sexual Diseases (Paraphilia) Using Method Dempster-Shafer. Journal of Intelligent Decision Support System (IDSS), 5(1),28–36. https://doi.org/10.35335/idss.v5i1.48.

[24] Michael, M. (2022). Expert System to Diagnose Bonsai Plant Pests with Certainty Factor Method. Journal of Intelligent Decision Support System (IDSS), 5(1), 20–27. https://doi.org/10.35335/idss.v5i1.77.

[25] Laia, T., & Hasugian, P. S. (2022). Expert System to Diagnose Eye Disease Due to Frequently Using Computer with Bayes Theorem Method. Journal of Intelligent Decision Support System (IDSS), 5(1), 1–9. https://doi.org/10.35335/idss.v5i1.46.

[26] Majumder, S., & Kar, S. (2018). Multi-criteria shortest path for rough graph. Journal of Ambient Intelligence and Humanized Computing, 9(6), 1835–1859. https://doi.org/10.1007/s12652-017-0601-6.

[27] Rupaidi. B, R. G. T. Utami, & M. E. P. Negara. (2020).” Coordination Pattern Between The Department Of Transportation And Banjarmasin Police Traffic Unit In Overcoming Congestion. Comparative Review”. Democracy In Digital Era: Law, Governance, Sosial And Economic Perspective In Asia, Australia And Dutch”. International Conference and Call Paper.

[28] Nurwan, at. All. (2021). Implementation of Dijkstra Algorithm and Welch-Powell Algorithm for Optimal Solution of Campus Bus Transportation. Jurnal Matematika MANTIK Vol. 7, No. 1, May 2021, pp. 31-40 ISSN: 2527-3159 (print) 2527-3167 (online). DOI : 10.15642/mantik.2021.7.1.31-40.

[29] Nagar, A., & Tawfik, H. (2007). A Multi-Criteria Based Approach to Prototyping Urban Road Networks. Issues in Informing Science and Information Technology, 4, 749–756. https://doi.org/10.28945/985.

[30] Rijalul Haqqi, Horas. SM Marpaung, M. S. (2017). Analisis Waktu Tempuh Kendaraan Bermotor Dengan Metode Kendaraan Bergerak. Jom FTEKNIK Fakultas Teknik: Universitas Riau., 2(34), 1–8.

[31] Oghanian, E., & Kebria, Z. S. (2017). The combination of TOPSIS method and Dijkstra’s algorithm in multi-attribute routing. Scientia Iranica, 24(5), 2540–2549. https://doi.org/10.24200/sci.2017.4390.

[32] Alem, I. E., Mijwil, M. M., Abdulqader, A. W., & Ismaeel, M. M. (2022). Flight schedule using Dijkstra’s algorithm with comparison of routes findings. International Journal of Electrical and Computer Engineering, 12(2), 1675–1682. https://doi.org/10.11591/ijece.v12i2.pp1675-1682.

[33] Ilveira, P., Teixeira, A. P., & Guedes Soares, C. (2019). AIS-based shipping routes using the Dijkstra algorithm. TransNav, 13(3), 565–571. https://doi.org/10.12716/1001.13.03.11.

[34] Ofallis, C. (2014). Add or Multiply? A Tutorial on Ranking and Choosing with Multiple Criteria. INFORMS Transactions on Education, 14(3), 109–119. https://doi.org/10.1287/ited.2013.0124.

[35] Idianto, T. (2021). Optimalisasi Metode Weighted-Sum Dijkstra Algorithm untuk Menentukan Rute Terbaik yang Sesuai dengan Kebutuhan. Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer Dan Teknologi Informasi, 7(1), 13. https://doi.org/10.24014/coreit.v7i1.11625.

[36] U, L., Kong, D., Shao, X., & Yan, X. (2018). A path planning and navigation control system design for driverless electric bus. IEEE Access, 6(c), 53960–53975. https://doi.org/10.1109/ACCESS.2018.2868339.

[37] http://rc.korlantas.polri.go.id:8900/eri2017/laprekappolda.php.

[38] https://www.pekanbaru.go.id/.

[39] Salem, I. E., Mijwil, M. M., Abdulqader, A. W., & Ismaeel, M. M. (2022). Flight schedule using Dijkstra’s algorithm with comparison of route findings. International Journal of Electrical and Computer Engineering, 12(2), 1675–1682. https://doi.org/10.11591/ijece.v12i2.pp1675-1682.

[40] Abdullah, N., & Hua, T. K. (2018). Weighted Methods of Multi-Criteria Via Dijkstra’s Algorithm in Network Graph For Less Congestion, Shorter Distance, and Time Travel in Road Traffic Network. NETWORK GRAPH FOR LESS CONGESTION, SHORTER DISTANCE, AND TIME. July. https://doi.org/10.5281/zenodo.1305010.

[41] Abdullah, N., & Hua, T. K. (2018). Weighted Methods of Multi-Criteria Via Dijkstra’s Algorithm in Network Graph for Less Congestion, Shorter Distance, and Time Travel in Road Traffic Network. NETWORK GRAPH FOR LESS CONGESTION, SHORTER DISTANCE, AND TIME. July. https://doi.org/10.5281/zenodo.1305010.

[42] Bunaen Chatrin, Pratiwi Hanna, & Riti Yosefina. (2022). Penerapan Algoritma Dijkstra Untuk Menentukan RuteTerpendek Dari Pusat Kota Surabaya Ke Tempat Bersejarah. Jurnal Teknologi Dan Sistem Informasi Bisnis, 4(1), 213–223.

[43] Chilukuri, B. R., Laval, J. a., Guin, A., Laval, J. a., Chilukuri, B. R., Technology, G. I. of, Transportation, G. D. of, Administration, F. H., (GTI-UTC), G. T. I. U. T. C., & Administration, R. and I. T. (2013). Freeway Travel-time Estimation and Forecasting. Transportation Research Record, 57p.

[44] http://www.utc.gatech.edu/sites/default/files/projects/reports/guin_laval__freeway_travel_time_estimation_0.pdf%0Ahttp://g92018.eos-intl.net/eLibSQL14_G92018_Documents/10-22.pdf%0Ahttps://trid.trb.org/view/1246696.

[45] Cutolo, A., De Nicola, C., Manzo, R., & Rarità, L. (2012). Optimal paths on urban networks using travelling times prevision. Modelling and Simulation in Engineering, 2012. https://doi.org/10.1155/2012/564168.

[46] Disser, Y., Müller-Hannemann, M., & Schnee, M. (2008). Multi-criteria shortest paths in time-dependent train networks. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 5038 LNCS, 347–361. https://doi.org/10.1007/978-3-540-68552-4_26.

[47] Hj, I. H., & Saputra, M. T. Y. (2021). Pengaruh Volume Lalu Lintas Terhadap Tingkat Kebisingan Disimpang Tiga Jalan Raya Bastiong. Clapeyron : Jurnal Ilmiah Teknik Sipil, L(2), 59–65. https://doi.org/10.33387/clapeyron.v2i2.3904.

[48] Kim, H., Kim, S., Hyoung, S., & Jang, K. (2013). Assessment of Travel Time Estimates based on Different Vehicle Speed Data : Spot Speed vs Sampled Journey Speed in South Korean expressways. Proceedings of the Eastern Asia Society for Transportation Studies, 9, 15. http://easts.info/online/proceedings/vol9/PDF/P334.pdf.

[49] Eklund, P.W., Kirkby, S. and Pollitt, S. \A dynamic multi-source Dijkstra's algorithm for vehicle routing", In Intelligent Information Systems, Australian and New Zealand Conference on, pp. 329-333, IEEE (1996).

[50] Peyer, S., Rautenbach, D. and Vygen, J. \A generalization of Dijkstra's shortest path algorithm with applications to VLSI routing", Journal of Discrete Algorithms, 7(4), pp. 377-390 (2009). doi:10.1016/j.jda.2007.08.003.

[51] Rodegerdts, L. A., Nevers, B., Robinson, B., Ringert, J., Koonce, P., Bansen, J., Nguyen, T., McGrill, J., Stewart, D., Suggett, J., & Neuman, T. (2004). Signalized intersections: informational guide (No. FHWA-HRT-04-091).

[52] P. D. P. Adi, A. Kitagawa, D. A. Prasetya and A. B. Setiawan, "A Performance of ES920LR LoRa for the Internet of Things: A Technology Review," 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT), Surabaya, Indonesia, 2021, pp. 1-7, doi: 10.1109/EIConCIT50028.2021.9431912.

[53] Fransiska Sisilia Mukti, Puput Dani Prasetyo Adi, Dwi Arman Prasetya, Volvo Sihombing, Nicodemus Rahanra, Kristia Yuliawan and Julianto Simatupang. “Integrating Cost-231 Multiwall Propagation and Adaptive Data Rate Method for Access Point Placement Recommendation”. International Journal of Advanced Computer Science and Applications (IJACSA) 12.4 (2021). http://dx.doi.org/10.14569/IJACSA.2021.0120494.

Embedding Models: A Comprehensive Review with Task-Oriented Assessment

Lahbib Ajallouda[1], Meriem Hassani Saissi[2], Ahmed Zellou[3]
Poly-Disciplinary Faculty of Es-Semara, Ibn Zohr University in Agadir, Morocco[1]
SPM-ENSIAS, Mohammed V University, Rabat, Morocco[2, 3]

***Abstract*—Sentence embedding is a very important technique in most natural language processing (NLP) tasks, such as answer generation, semantic similarity detection, text classification and information retrieval. This technique aims to transform the semantic meaning of a sentence into a fixed-dimensional vector, allowing machines to understand human language. Sentence embedding has moved in recent years from simple word vector averaging methods to the development of more sophisticated models, particularly those based on transformer structures such as the BERT model and its variants. However, systematic reviews that critical, analyze and compare the performance of these models are still limited, particularly the selection of the appropriate embedding model for a specific NLP task. This study aims to address this gap by a comprehensive review for sentence embedding models and a systematic evaluation of their performance on NLP tasks, such as semantic similarity, clustering, and retrieval. The study enabled us to identify the appropriate embedding model for each task, identify the main challenges faced by embedding models, and propose effective solutions to improve the performance and efficiency of sentence embedding.**

***Keywords*—Natural language processing; sentence embedding models; transformer models; embedding models challenges**

I. Introduction

Embedding is one of the most techniques used in natural language processing applications [1] [2]. The performance of these applications is affected by the quality of the fixed-dimensional vectors generated [3]. Several studies have underscored the importance of developing sentence embedding models [4], [5], especially with the capabilities offered by deep neural networks and transformers. NLP tasks require a deeper understanding of texts, especially understanding long texts [6], [7]. Therefore, choosing an appropriate embedding model is essential for a task. Recent frameworks such as DSPy [8] underscore the need for more sophisticated and context-aware sentence representations that are better suited to specific NLP tasks. Therefore, there is a need to study and analyze sentence embedding techniques, from statistical models based on word vector averages to modern models based on deep structures such as BERT [9] and its variants. In this context, we conducted a comprehensive literature review of the most sentence embedding models, from simple aggregation techniques to transformer-based models.

The main objective of our paper is to identify the most appropriate embedding model for a specific task. To achieve this, our paper first presents the most important embedding models, their characteristics, the challenges that affect their effectiveness, and solutions that can help overcome them. Second, it evaluates the performance of these models on three natural language processing tasks [10], including semantic similarity, clustering, and retrieval, to identify the most effective models for each task. The methodology used in our paper is based on answering three main research questions:

- RQ 1: What are the main categories of sentence embedding models proposed in the literature?
- RQ 2: What are the strengths, limitations, and challenges of each model?
- RQ 3: Which model is appropriate for each NLP task?

We reviewed and analyzed scientific papers published in peer-reviewed sources (ArXiv, Scopus, Dblp, IEEE) between 2015 and 2025. The results of this study represent an opportunity for researchers to build an overview of the techniques used in sentence embedding, to advance this area.

The rest of the paper will be presented as follows. Section II presents related work on sentence embedding. Section III discusses categories of sentence embedding models, highlighting their advantages and challenges. Section IV is devoted to an empirical study that aims to evaluate these models across NLP tasks. Section V presents and discusses the results of the study. Section VI concludes the paper with future research directions.

II. Related Work

In this section, we first highlight the importance of embedding models for natural language processing tasks and applications. We then trace the evolution of sentence embedding models. Finally, we summarize the scope and limitations of previous reviews to underline the relevance of the present study.

A. Sentence Embeddings in NLP Tasks

Embedding models have attracted growing interest in NLP field, as they demonstrate strong effectiveness in a variety of tasks, such as retrieval, clustering, summarization, and semantic similarity [11]. The development of embedding techniques has positively impacted this area [12]. Recent research has demonstrated that their use significantly improves the performance of NLP applications. For example, the performance of text classification and clustering has improved by pre-trained embeddings [13]. Performance in dense retrieval across multiple datasets has improved by employing contrastive learning models [14]. Similarly, extractive summarization models have improved by enabling the selection of more semantically rich sentences compared to standard baselines

[15]. Clustering studies also confirm that embeddings derived from pre-trained large language models capture accurate semantic relationships [16]. These results confirm that exploiting sentence embeddings for a specific task is essential for improving the performance of NLP tasks and applications.

B. Sentence Embedding Evolution

A text embedding is a vector representation of text (word, sentence, or paragraph), where similar texts appear close together in the embedding space [5]. These representations encode the syntactic and semantic properties of linguistic elements. Classic representations have relied on the bag-of-words model or TF-IDF weighting [17], where each word is represented by a high-dimensional vector reflecting its frequency in the document (Formula 1). For a word w in a document d, the TF-IDF vector is defined as:

$$\vec{W} = (TFIDF(w,d_1), TFIDF(w,d_2), \dots, TFIDF(w,d_N)) \quad (1)$$

Where $TF-IDF(w,d)$ is calculated by Formula 2.

$$TF-IDF(w,d) = \frac{f_{w,d}}{\sum_{w' \epsilon d} f_{w',d}} \times log\left(\frac{N}{1+|\{d \epsilon D: w \epsilon d\}|}\right) \quad (2)$$

Where:

- $f_{w,d}$: Frequency of the word w in document d.
- N: Number of documents in corpus D.
- $|\{d \epsilon D: w \epsilon d\}|$: Number of documents containing the word w.

However, these models suffer from several limitations, including their high dimensionality and their inability to capture semantic similarity between terms.

Word embedding, through models such as Word2Vec [18] and GloVe [19], has enabled the representation of words using low-dimensional vectors while considering relevant semantic contexts. On the other hand, the same vector is assigned to a word regardless of its context. Therefore, these models ignore the context in which the word appears, making them unsuitable for modeling sentences or paragraphs. Fig. 1 provides an example of this challenge.

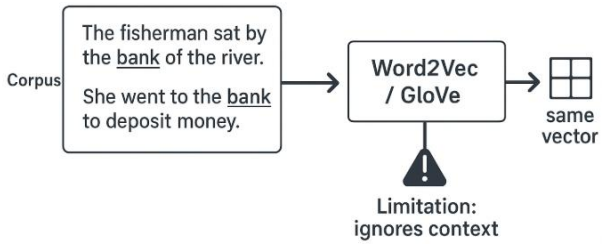


Fig. 1. Contextual limitation of Word2Vec and GloVe.

To address the need to model text units containing more than one word, sentence embeddings were developed that aim to represent each sentence as a fixed-dimensional dense vector.

Sentence embedding techniques are classified into two main categories [20]. The first includes statistical techniques, such as Doc2Vec [21] and SIF [22]. These techniques are computationally inexpensive. However, they remain limited because they only capture a portion of the semantic context. The second set consists of deep neural techniques, which rely on advanced encoding models. These include InferSent [23] and Universal Sentence Encoder [24], Sentence-BERT [25] and SimCSE [26]. These models use transformers to contextually model complex relationships between words and sentences. They therefore provide more expressive vector representations suitable for advanced NLP tasks.

C. Comprehensive Review of Sentence Embedding

Several research papers have studied and reviewed proposed sentence embedding models, offering different perspectives on their classification, evaluation, and application. The work presented in [27] is one of the first general reviews of sentence representations. This study was limited to unsupervised neural approaches. Other reviews such as [23], [28] have analyzed and compared sentence representation learning, proposed unified evaluation metrics such as SentEval [29], and focused on classical RNN and CNN encoders. In contrast, studies, such as [24], [25], have highlighted the effectiveness of transformer learning using pre-trained models such as BERT and its variants. Finally, recent published studies [26], [30] have focused on the importance of contrastive learning and pre-trained models. Despite the comprehensiveness of these reviews, the development of embedding models across NLP tasks, and the difficulty of identifying the appropriate model for each task, is a challenge that motivates a comprehensive and updated review that considers the importance of models to the nature of NLP tasks and their applications.

III. Sentence Embedding Models

This section introduces two main categories of sentence embedding models. Statistical sentence embedding models represent sentences using aggregated word-level statistics, while transformer-based sentence embedding models represent sentences using pre-trained contextual language models. This classification highlights the shift from shallow frequency-based representations to neural structures that better encode semantic and syntactic relationships.

A. Statistical Sentence Embedding Techniques

Statistical methods for sentence embedding rely on simple aggregation techniques, such as averaging or weighted summation of word embeddings, or distributional models. Although computationally efficient and easy to implement, these methods provide an approximate semantic representation of a sentence, ignoring word order and syntactic structures.

1) Unweighted averaging: Unweighted averaging is one of the most techniques widely used to generate sentence embeddings. The method consists of representing a sentence by computing the arithmetic mean of the vector representations of its constituent words. According to study [18], a sentence embedding is defined by Formula 3.

$$v_{sentence} = \frac{1}{n}\sum_{i=1}^{n} v_{w_i} \quad (3)$$

Where:

- n: Number of words in the sentence.
- v_{w_i} : The embedding representation of word w_i.

Despite its simplicity, it has been shown to capture some of the semantic meaning of vectors. However, its limitations include ignoring the significance of words in a sentence. It treats all words equally, including frequent terms.

2) TF-IDF weighted averaging: The TF-IDF weighted averaging incorporates word importance by exploiting the TF-IDF coefficient. This allows us to overcome the challenge of treating all words equally and reduce the influence of common words. Formally, according to [22], the sentence embeddings are obtained by calculating a weighted average of word vectors, where the weights are determined by TF-IDF coefficient (Formula 4).

$$V_{sentence} = \frac{\sum_{i=1}^{n} TF\text{-}IDF(w_i) \cdot v_{w_i}}{\sum_{i=1}^{n} TF\text{-}IDF(w_i)} \quad (4)$$

Where:

- n: Number of words in the sentence.
- v_{w_i} : The embedding representation of word w_i.
- $TF - IDF(w_i)$: Weight of word w_i, is calculated by Formula 2.

This model generates better semantic representations than unweighted averaging. However, it still ignores the syntactic information and sequential order of words, which may limit its effectiveness in more complex NLP tasks.

3) Max and min pooling models: Max and min pooling is a statistical model for generating sentence embeddings by selecting the maximum (or minimum) value of all word embeddings in a sentence. This approach captures the most prominent features present in any word vector, making it robust to noise. According to [23], the sentence embedding can be generated using either max-pooling or min-pooling as defined by Formulas (5) and (6).

$$V_{sentence}^{max} = \left[\max_{i=1..n}(v_{w_i}^{(1)}), \max_{i=1..n}(v_{w_i}^{(2)}), \dots \max_{i=1..n}(v_{w_i}^{(d)})\right] \quad (5)$$

$$V_{sentence}^{min} = \left[\min_{i=1..n}(v_{w_i}^{(1)}), \min_{i=1..n}(v_{w_i}^{(2)}), \dots \min_{i=1..n}(v_{w_i}^{(d)})\right] \quad (6)$$

Where:

- n: Number of words in the sentence.
- $v_{w_i}^{(k)}$: the value of the k-th dimension of word w_i (each of dimension d).

However, it still ignores syntactic information and may oversimplify the semantic representation. The semantic structure of sentences may be overly simplified.

4) Smooth inverse frequency model: The Smoothed Inverse Frequency (SIF) [22] model reduces the influence of common words on sentence embeddings by weighting word embeddings inversely with their frequency in a corpus. To generate sentence embeddings, SIF calculates a weighted average (Formula 7) and removes the first principal component that defines common directions in the embedding space to produce more discriminative sentence representations (Formula 8).

$$V_{sentence} = \frac{1}{n}\sum_{i=1}^{n} \frac{\alpha}{\alpha + p(w_i)} \cdot v_{w_i} \quad (7)$$

Where:

- n: Number of words in the sentence.
- v_{w_i} : Embedding representation of word w_i.
- $p(w_i)$: Frequency of w_i in the corpus.
- α: Smoothing parameter (default 10^{-3})

To remove the first principal component F that defines the common directions in the embedding space, we need to calculate $FF^T V_{sentence}$ the Projection of $V_{sentence}$ onto F ($F^T V_{sentence}$ is the scalar product between F and $V_{sentence}$).

$$V_{Sentence}^{SIF} = V_{sentence} - FF^T V_{sentence} \quad (8)$$

Although SIF is efficient, and robust, but its reliance on frequency statistics, principal component analysis calculations, and inability to capture syntax or context limits its effectiveness for more complex NLP tasks.

B. Transformer-Based Sentence Embedding Models

Sentence embedding using transformers has contributed to generating contextual representations that capture the semantic meaning of sentences and reducing the challenges of statistical models. Bidirectional Encoder from Transformers (BERT) [9] is the first transformer used for sentence embedding. Other transformers have followed, improving the performance of the sentence embedding task. In this part, we provide a comprehensive overview of the transformer-based sentence embedding models, based on the classification shown in Fig. 2.

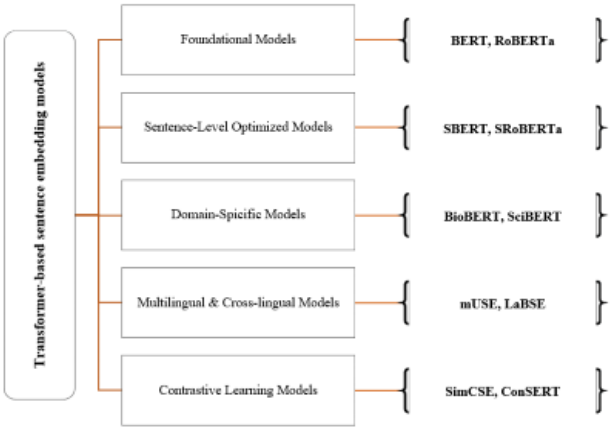


Fig. 2. Taxonomy of transformer-based sentence embedding models.

1) Foundational models: BERT is the first transformer model for sentence embedding. Unlike statistical approaches, BERT relies on self-attention mechanisms rather than recurrent or convolutional networks. It has bidirectional encoding, which allows it to read sequences of tokens in both directions. It uses multiple transformer layers (12 in BERT-base and 24 in BERT-large), each consisting of multi-head self-attention networks and feed-forward networks. Its pre-training combines masked language modeling (MLM), where random tokens are predicted from context, and next-sentence prediction (NSP), which

captures sentence-level relationships. Fig. 3 illustrates the process by which BERT produces these embeddings. These features make BERT more effective to learn deep contextual representations.

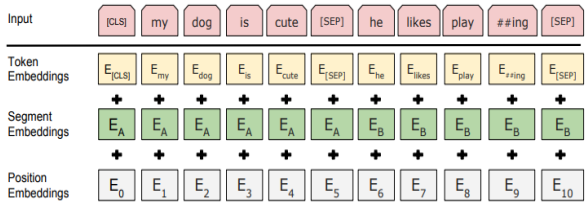


Fig. 3. The BERT sentence embedding process, reproduced from study [9].

However, the computational cost of pretraining, biases in pretraining data, and sequence length constraints limit its performance to represent long documents. This challenge has prompted the development of BERT variants such as RoBERT [31] that uses the same transformer encoder architecture as BERT, but with improvements in the pre-training phase. It Eliminates the NSP task, employs dynamic masking in the MLM objective, and trains on larger datasets (over 160 GB). This allows it to benefit from longer training schedules and higher learning, enhancing the quality of its contextual representations.

Despite their effective performance, Foundational models face several challenges. The large training dataset makes them computationally expensive, and they are also affected by biases present in pre-training datasets. Furthermore, the input length limit of 512 tokens limits their ability to represent longer sequences.

2) Sentence-level optimized models: To overcome the challenges of foundational models, sentence-level optimized models, such as SBERT and SRoBERTa [25], rely on Siamese and triplet network architectures, which improve the accuracy of transformer encoders to generate semantically meaningful sentence representations. They also reduce computational cost on semantic text similarity (STS) tasks. The time to find the most similar pair of sentences in a set of 10,000 sentences is reduced from 65 hours using BERT to about 5 seconds using SBERT. Using a Siamese and triplet network also embeds sentences into higher-dimensional semantic spaces to better capture hierarchical information.

TABLE I. KEY DISTINCTIONS BETWEEN SBERT AND SROBERTA

Criterion	SBERT	SRoBERTa
Encoder	BERT or RoBERTa	RoBERTa (with enhanced pretraining)
Training Data	STS/NLI datasets	Larger and more diverse datasets, often domain-adapted
Specialized Tasks	Semantic similarity	Semantic similarity, clustering
Performance	Strong for semantic tasks	Higher performance, especially in specific domain

Table I shows the key differences between SBERT and SRoBERTa. In SBERT, the input sentence is encoded using BERT or RoBERTa, followed by a pooling layer that produces a fixed-size embedding, which is then applied to semantic similarity tasks. In contrast, SRoBERTa uses the RoBERTa encoder and incorporates structured training strategies, such as variance and classification losses, to produce more accurate embeddings. These embeddings are particularly effective for semantic search and clustering, and therefore SRoBERTa outperforms SBERT.

The performance of SBERT and SRoBERTa on Cross-lingual data is low, compared to their results on English data. Therefore, their performance depends on the language of their training data (English) which makes their generalization to complex languages difficult. Therefore, multilingual and cross-linguistic sentence embedding models are important to achieve robust performance on multilingual applications.

3) Multilingual and cross-linguistic models: Semantic similarity Detection between sentences across different languages has been a challenge for monolingual sentence embedding models. In contrast, multilingual and cross-linguistic models, such as the Multilingual Universal Sentence Encoder (mUSE) [32] and Language-Independent BERT Sentence Embeddings (LaBSE) [33], achieve minimal linguistic bias, leading to better performance.

The mUSE model uses a universal sentence encoder and is trained on a set of translation pairs in multilingual datasets. mUSE supports more than 16 universal languages, and achieves a semantic similarity and effective cross-language retrieval. However, its performance deteriorates on untrained languages, this limit its application in global contexts. Unlike mUSE, LaBSE is a language-agnostic model trained on over 110 languages using a dual-encoder architecture and translation ranking objectives. LaBSE achieved state-of-the-art results on benchmarks such as Tatoeba [34] and cross-linguistic STS [35]. It demonstrated its ability to encode semantically equivalent sentences from different languages, making it one of the most robust models for multilingual applications.

Overall, these models have achieved the transition from English-centric sentence embeddings to multilingual sentence embeddings. Despite this contribution, these models still face significant challenges. Their reliance on massive parallel datasets and complex training pipelines limits their adaptability to resource-limited languages. They also rely on translation-based alignment objectives, which can introduce biases and limit the generalization of sentence representations in monolingual or domain-specific. To overcome these challenges, contrastive learning models have emerged, which aim to learn sentence embeddings by exploiting self-supervised objectives.

4) Contrastive learning models: To overcome the challenges faced by multilingual sentence embedding models, contrastive learning models has been adopted as a paradigm for learning universal sentence representations. these models aim to approximate semantically similar sentences in the embedding space while separating unrelated sentences, without the need for massive parallel datasets. SimCSE [26] and ConSERT [36] are the most widely used contrastive learning models for sentence embedding.

SimCSE relies on two different representations of the same sentence by applying contrastive dropout masks within a pre-

trained model such as BERT or RoBERTa. These representations are considered as positive pairs, while the other sentences are treated as negative samples. By the contrastive loss optimization (InfoNCE) [37], SimCSE is learn to align semantically consistent representations without the need for external supervision. The ConSERT sentence embedding model relies on contrastive learning to efficiently fine-tune BERT, leveraging positive pair construction through strategies including word reordering, token masking, and embedding dimension masking. Sentence representations are obtained by aggregating mean pooling instead of CLS. Which makes it able to adapt in low-resource contexts.

However, some challenges are remained. These models require fine-tuning of hyperparameters (dropout rates, and augmentation strategies), making their performance unstable across domains. Their performance is also affected in resources lacking sufficient negative samples due to their reliance on aggregated negatives. Furthermore, these models are largely trained on English corpora, offering limited cross-language generalization without extensive retraining.

IV. Empirical Study

This section presents the evaluation process adopted in our study. Fig. 4 shows the most important stages of this process. We introduce the benchmark datasets commonly used to evaluate the performance of sentence embedding models. Next, we present the evaluation metrics used to measure the performance of NLP tasks such as semantic similarity, clustering, retrieval information, and classification. We present empirical results on the performance of these models. Finally, we discuss the results, identifying the appropriate models for each NLP task.

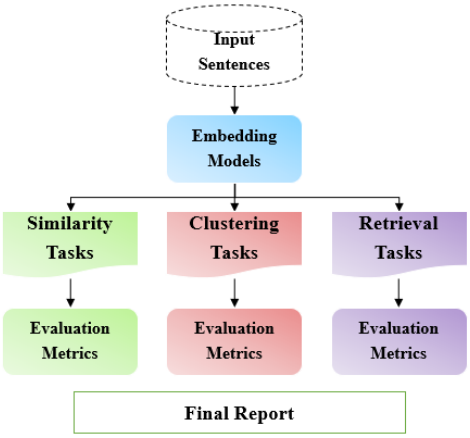


Fig. 4. Process of the evaluation sentence embedding models.

In the first stage, sentences are received by the embedding model to be evaluated, which converts them into numerical vectors in the second stage. These vectors are tested on three main tasks in the third stage. Similarity tasks to measure the degree of meanings similarity, clustering tasks to test the ability to cluster similar sentences, and retrieval tasks to evaluate the effectiveness of sentence retrieval. For each task, appropriate evaluation metrics are adopted to reflect the model's accuracy and performance. Finally, the results are analyzed in a comprehensive report that allows for model comparison.

A. Evaluation Datasets

The evaluation of sentence embedding models relies on datasets that capture task requirements. Semantic text similarity (STS) datasets, such as STS-Benchmark [35] and SICK-R [38], are used because they allow for easy interpretation of the semantic quality scores of embeddings. However, its small size and focus on a specific domain limit its generalizability. Therefore, Natural Language Inference (NLI) datasets, including SNLI [39] and MultiNLI [40], are also commonly used to evaluate embedding techniques. Despite its large size and diversity of sentence pairs, it can lead to embedding without a real understanding of the meaning. Datasets such as MRPC [41] and QQP [42] are used to measure the ability of embeddings to capture paraphrasing and lexical diversity. For retrieval and clustering tasks, datasets such as MS MARCO [43], BEIR [44], and StackExchange [45] are used. Their challenges lie in high variability and data noisy.

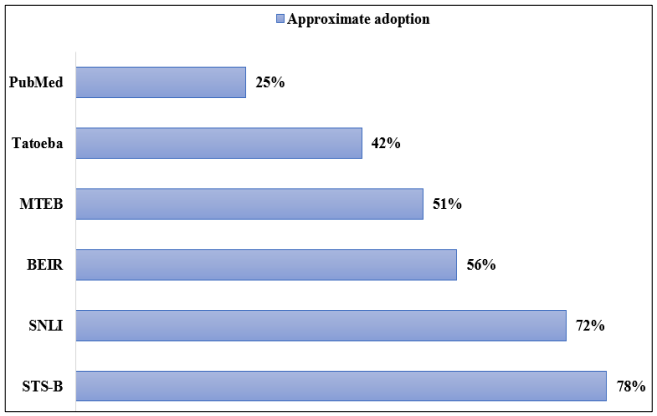


Fig. 5. Approximate adoption rates of training and evaluation datasets in sentence embedding research.

The multilingual datasets, including XNLI [46], Tatoeba [47], and the MTEB [48] multilingual path, are essential for evaluating models across different languages. However, they remain limited in low-resource languages. Fig. 5 shows the approximate adoption rates of datasets in sentence embedding research. Table II compares the most important datasets used for evaluation, in terms of task, size, language, and challenges.

Domain-specific models such as BioBERT, SciBERT instead rely on specialized datasets such as the PubMed biomedical dataset or scientific papers from Semantic Scholar.

B. Evaluation Metrics

To measure the performance of sentence embedding models, the NLP tasks identify the appropriate metric. For semantic textual similarity (STS) tasks, correlation metrics such as Pearson correlation coefficient and Spearman correlation coefficient [49] are used. Cosine similarity [50] is used to measure how closely embeddings converge in vector space, making it suitable for tasks such as clustering and paraphrase detection.

TABLE II. COMPARISON OF DATASETS USED TO TRAIN AND EVALUATE SENTENCE EMBEDDING MODELS

Dataset	Task	Size	Language	Challenges
STS-B	Semantic Textual Similarity	8,600 pairs	English	Small size, domain-limited
SICK-R	Semantic Textual Similarity + Relatedness	10,000 pairs	English	Domain-limited, relatively small
SNLI	Natural Language Inference	570,000 pairs	English	Annotation in English-only
MultiNLI	Natural Language Inference	430,000 pairs	English (multiple genres)	Still English-only, annotation bias
MRPC	Paraphrase Identification	5,800 pairs	English	Small dataset, domain bias (news)
QQP	Paraphrase Identification	400,000 pairs	English	Noise, duplicates, biased toward question style
MS MARCO	Passage Retrieval / Ranking	1M passages, 100k queries	English	Noisy labels, domain-specific (web queries)
StackExchange	Clustering & Retrieval	100k+ QA pairs	English	Annotation ambiguity , Duplicate questions
XNLI	Cross-lingual NLI	750k (15 languages)	Multilingual	Limited low-resource language support
Tatoeba	Cross-lingual Sentence Similarity	1,000+ pairs per language	100+ languages	Uneven quality across languages
MTEB	Multi-task, multi-lingual (STS, retrieval, classification, clustering)	8M examples across 50 datasets	Multilingual (100+ languages)	Complex to run, requires large compute

Also, for clustering tasks, Normalized Mutual Information (NMI), Silhouette score, and Adjusted Rand Index (ARI) [51], can be used to evaluate how well embeddings cluster semantically similar sentences. For classification and natural language inference (NLI) tasks, Accuracy and F1-Score [52] are the most appropriate metrics. For information retrieval and classification tasks, metrics such as mean average precision (MAP) and mean reciprocal rank (MRR) [53] are preferred. Table III presents the most important metrics for evaluating sentence embedding models for each NLP task.

TABLE III. EVALUATION METRICS FOR SENTENCE EMBEDDING MODELS ON NLP TASKS

Metric	STS	Clustering	Paraphrase	Retrieval	NLI	Classification
Cosine Similarity	X	X	X			
Pearson Correlation	X					
Spearman Correlation	X					
Accuracy					X	X
F1-Score			X		X	X
Mean Average Precision (MAP)				X		
Mean Reciprocal Rank (MRR)				X		
Normalized Mutual Information (NMI)		X				
Adjusted Rand Index (ARI)		X				
Silhouette Score		X				

Analysis of sentence embedding research shows that most sentence embedding models prefer evaluation based on semantic textual similarity tasks, while evaluation tasks such as information retrieval and classification are neglected. Fig. 6 shows the approximate adoption rates of evaluation metrics in sentence embedding research.

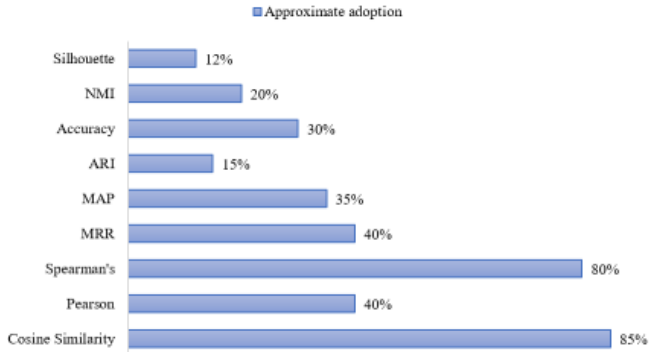


Fig. 6. Approximate adoption rates of evaluation metrics in sentence embedding research.

C. Evaluation protocol

1) Statistical sentence embedding techniques: To evaluate statistical sentence embedding techniques, popular datasets and evaluation metrics were used. The STS-Benchmark dataset and Spearman and Pearson correlations were used to measure the effectiveness of embedding on semantic textual similarity (STS). For sentiment classification tasks, the Stanford Sentiment Treebank (SST-2) dataset [54] and the classification accuracy metric were used. Additionally, the clustering quality was measured on sentence embeddings derived from the STS-B corpus using the silhouette score.

According to the evaluation results presented in Fig. 7, the performance of unweighted averaging was affected by uniform treatment of words, especially in semantic similarity tasks. The

results of TF-IDF weighted averaging also make it more suitable for classification tasks. While the performance of max/min pooling, which oversimplifies the semantic of a sentence, remains inconsistent.

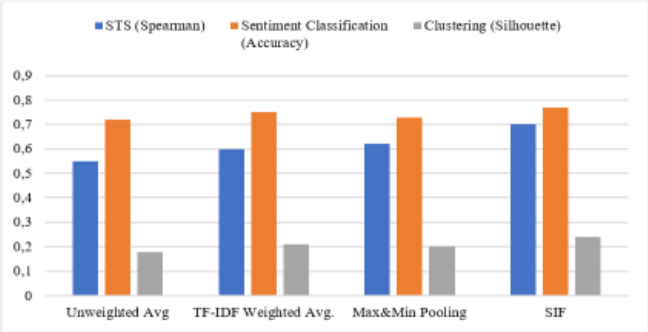


Fig. 7. Performance of statistical sentence embedding techniques across STS, sentiment classification, and clustering.

Conversely, the results demonstrate that SIF outperforms other methods by removing the most common directions in the embedding space, especially in semantic similarity and clustering tasks.

2) Transformer-based models: To ensure a comprehensive evaluation of sentence embedding models, three NLP tasks were adopted. Semantic similarity, clustering, and retrieval. This evaluation requires specific datasets and evaluation metrics for each task. For semantic similarity task, STS-B and Tatoeba were used. STS-B is a widely used for semantic similarity in English, while Tatoeba extends the evaluation to include multilingual sentence pairs. The performance was measured using Pearson and Spearman coefficients. In contrast, for the clustering task, MTEB and STS Clustering were adopted, which test the ability of embeddings to cluster semantically similar texts. Cosine similarity and silhouette score are used to measure the performance. Retrieval task was evaluated using BEIR and MIRACL datasets. BEIR provides a unified framework for information retrieval, while MIRACL focuses on cross-linguistic retrieval performance. MAP and MRR, which reflect the importance of retrieval, were used to measure the performance. Table IV presents the performance of the sentence embedding models on the semantic similarity task evaluation in STS-B and Tatoeba datasets.

The results showed modest performance of BERT and RoberTa, compared to the effectiveness of SBERT and SRoBERTa, on semantic similarity tasks. The multilingual models (mUSE and LabSE) also demonstrated effective performance, particularly on Tatoeba, confirming their performance on cross-linguistic representations. In addition to their computational efficiency, the contrastive learning models (SimCSE and ConSERT) achieved competitive results, making them attractive for light applications. Overall, these results suggest that domain, resource constraints, and language requirements determine the appropriate model.

The performance of sentence embedding models on the clustering task was evaluated using the MTEB Clustering and STS Clustering datasets as a practical benchmark for document clustering. Performance is measured using cosine similarity, which measures the convergence of embeddings within clusters, and the Silhouette score, which reflects cohesion and separation within a cluster. Table V presents the performance results of these models for each dataset.

TABLE IV. PERFORMANCE OF THE TRANSFORMER-BASED MODELS ON THE SEMANTIC SIMILARITY TASK

Model	STS-B		Tatoeba	
	Pearson	Spearman	Pearson	Spearman
BERT	0.63	0.60	0.42	0.40
RoBERTa	0.66	0.65	0.45	0.42
SBERT	**0.85**	**0.81**	0.72	0.70
SRoBERTa	**0.86**	**0.82**	0.71	0.68
BioBERT	0.60	0.55	0.40	0.43
SciBERT	0.70	0.62	0.50	0.46
mUSE	0.76	0.72	**0.85**	**0.82**
LaBSE	0.78	0.75	**0.88**	**0.86**
SimCSE	0.82	0.79	0.78	0.75
ConSERT	0.80	0.76	0.76	0.73

TABLE V. PERFORMANCE OF THE TRANSFORMER-BASED MODELS ON THE CLUSTERING TASK

Model	MTEB Clustering		STS Clustering	
	Cosine	Silhouette	Cosine	Silhouette
BERT	0.38	0.14	0.42	0.15
RoBERTa	0.40	0.16	0.44	0.20
SBERT	0.58	0.27	0.60	0.30
SRoBERTa	0.62	0.31	0.63	0.30
BioBERT	0.45	0.20	0.50	0.20
SciBERT	0.48	0.23	0.54	0.23
mUSE	0.58	0.25	0.60	0.25
LaBSE	**0.62**	**0.29**	0.62	0.28
SimCSE	**0.64**	**0.32**	**0.67**	**0.31**
ConSERT	0.61	0.28	**0.65**	**0.29**

The clustering results demonstrate the superior performance of the SimCSE and ConSERT models, confirming the effectiveness of contrastive learning to produce embeddings that enable the formation of coherent clusters. However, the performance of domain-specific models, such as BioBERT and SciBERT, is affected by the specialization of their training data and thus unsuitable for clustering tasks. In contrast, the multilingual models (mUSE and LabSE) provide competitive results. LaBSE achieves strong performance in multilingual clustering, making it suitable for clustering multilingual documents.

The performance of sentence embedding models on the retrieval task was evaluated using the BEIR and MIRACL benchmarks, which serve as widely adopted frameworks for testing retrieval effectiveness across both monolingual and multilingual settings. Evaluation was carried out using Mean

Average Precision (MAP), which measures the overall ranking quality, and Mean Reciprocal Rank (MRR), which reflects the ability of models to return relevant results at top ranks. Table VI presents the results of all models across both datasets.

TABLE VI. PERFORMANCE OF THE TRANSFORMER-BASED MODELS ON THE RETRIEVAL TASK

Model	BEIR		MIRACL	
	MAP	MRR	MAP	MRR
BERT	0.27	0.30	0.21	0.25
RoBERTa	0.30	0.32	0.23	0.26
SBERT	0.41	0.45	0.35	0.38
SRoBERTa	0.44	0.46	0.40	0.42
BioBERT	0.33	0.37	0.20	0.23
SciBERT	0.35	0.40	0.22	0.24
mUSE	0.32	0.35	**0.42**	**0.46**
LaBSE	0.34	0.39	**0.45**	**0.48**
SimCSE	**0.48**	**0.52**	0.41	0.45
ConSERT	**0.45**	**0.50**	0.38	0.41

Evaluation results showed that baseline models, such as BERT and RoberTa, underperformed on the retrieval task. In contrast, fine-tuning the semantic similarity tasks in SBERT and SRoBERTa improved this performance. Domain-specific models, such as BioBERT and SciBERT, faced challenges in the cross-domain. Contrast learning models (SimCSE and ConSERT) produced embeddings that improved the retrieval task. In contrast, cross-linguistic models (mUSE and LabSE) outperformed only on the MIRACL dataset, where cross-linguistic coverage is essential, but their performance lagged behind the English models in BEIR.

V. DISCUSSION

The evaluation of sentence embedding models focused on quantitative results for natural language processing tasks, including semantic similarity, clustering, and retrieval. While these results provide a numerical comparison of performance, a deeper analysis requires discussion and interpretation that considers practical aspects, such as training costs, model size, generalizability, and cross-linguistic applicability. This section interprets the results and links them to the research questions. This provides a consistent understanding of how each result contributes to answering these questions. This will enable researchers to use embedding models effectively in NLP tasks.

- RQ1: What families of sentence embeddings exist?

Regardless of whether the models are supervised or unsupervised, our results identify two main categories:

Statistical models such as Bag-of-Words, TF-IDF, and SIF, which are fast, interpretable, and easy to implement but fail to capture word order and deeper semantic relations. Statistical models are still inexpensive, but their performance on semantic tasks is low.

Transformer-based models such as BERT, SBERT, SimCSE and LaBSE, which produce contextualized and semantically rich representations. Transformer-based models, on the other hand, achieve significantly higher scores on NLP tasks (semantic similarity, clustering, and retrieval).

- RQ2: What are their strengths and limitations?

Statistical models are simple and easy to interpret, making them suitable for fields with limited resources. However, their representation neglects word order and context, limiting their semantic performance. A comparative analysis showed differences between the four statistical models in simplicity, interpretation, and representation. The unweighted average model provides a straightforward baseline with minimal computational cost, but it often fails to capture semantic relationships. The weighted average TF-IDF model focuses on meaningful words in a sentence, making it more effective for classification tasks. Max-min pooling provides richer embeddings by considering extreme feature values, which can better reflect sentence variance. Finally, the SIF model emerges as a more accurate approach, balancing simplicity and semantic accuracy by reducing the dominance of frequent words and processing meaning at the sentence level.

Transformer-based sentence embedding models also have strengths and weaknesses. Foundational models, such as BERT and RoBERTa, have formed the basis of most models, but they are not optimized sentence level and lack cross-linguistic support. Optimized models, such as SBERT and SRoBERTa, have improved performance on NLP tasks, but they require fine-tuning specific to the task and remain mostly English-focused. While specialized models, such as BioBERT and SciBERT, are effective at embedding biomedical and scientific terms, their performance declines outside of specialized domains. Contrast learning models, such as SimCSE and ConSERT, produce embeddings that achieve high performance on similarity and clustering through simple and efficient contrastive objectives, although they are highly sensitive to training strategies and require large datasets. The main strength of multilingual models, such as mUSE and LabSE, is their ability to exploit common embedding spaces across more than 100 languages, particularly for cross-linguistic retrieval. However, their performance declines in low-resource languages and requires intensive training resources.

Overall, the analysis results indicate that statistical embedding models are more suitable for simple applications, while transformer-based models have become the preferred choice for most NLP tasks, especially those requiring precise semantic analysis. Future research could explore hybrid strategies that combine the efficiency of statistical models with the semantic depth of transformers.

- RQ3: What model is appropriate for each NLP task?

The evaluation results, presented in Tables IV, V, and VI, demonstrate the appropriate sentence embedding model for each NLP task. For semantic similarity tasks, SRoBERTa and SBERT achieved the highest correlation rates on the STS dataset, thus capturing subtle semantic nuances. While LaBSE achieved the highest correlation rates on the Tatoeba dataset, making it more suitable for multilingual scenarios. For clustering tasks, SimCSE outperformed the other models, achieving sentence clustering without being affected by

multilingualism. For retrieval tasks, SimCSE and ConSERT performed best, followed by SRoBERTa, making them most effective for monolingual retrieval tasks. Conversely, LaBSE outperformed on the MIRACL dataset, making it more suitable for multilingual retrieval tasks. Table VII presents the classification of NLP tasks according to appropriate sentence embedding model.

TABLE VII. CLASSIFICATION OF NLP TASKS ACCORDING TO BEST SENTENCE EMBEDDING MODELS

Task	Best Models	Description
Retrieval	SimCSE, ConSERT, SBERT	Contrastive models (SimCSE, ConSERT) are strong baselines, while SBERT remain very competitive.
Clustering	SimCSE, ConSERT	Contrastive models generate well-separated embeddings for clustering.
Semantic Similarity	SBERT, SRoBERTa	SBERT is traditionally strongest for similarity tasks
Multilingual tasks	LaBSE	LaBSE is outperformed for Multilingual tasks

In general, SBERT and SRoBERTa are suitable for semantic similarity, SimCSE and ConSERT are effective for retrieval and clustering, and LaBSE is best suited for multilingual tasks.

VI. CONCLUSION

The study presented in our paper combines a review and empirical assessment of sentence embedding models to address the challenge of selecting the most appropriate model for each NLP task. The study answers three main research questions. First, (RQ1), there are two main categories of embeddings: statistical models such as unweighted average, weighted average TF-IDF, Max-min pooling and SIF, which remain simple and interpretable, and transformer-based models such as BERT, RoBERTa, SBERT, BioBERT, LabSE and SimCSE, which exploit contextual representations. Second, (RQ2), we analyzed their strengths and weaknesses, while lightweight statistical models appear to be incapable of semantic representations. Transformer-based models capture rich contextual information but require significant computational resources and may face challenges in domain transfer. Finally, (RQ4), we determined the appropriate model for three specific NLP tasks. Our empirical study showed that SBERT and SRoBERTa produce embeddings that are well-suited for semantic similarity tasks. In contrast, the embedding generated by the SimCSE model performs best for clustering tasks, while the embeddings generated by the SimCSE, ConSERT, and SRoBERTa models are well-suited for retrieval tasks. In general, the selection of an appropriate model depends on the task requirements, available resources, and language context. For low-resource or interpretable applications, statistical embeddings remain appropriate. For semantically rich applications, despite their computational cost, transformer models are highly efficient. In contrast, multilingual applications can benefit from models such as LaBSE. Future work should explore hybrid strategies that combine efficiency and semantic depth, and expand the scope of evaluations to include multimodal embeddings to meet the growing needs of NLP systems.

REFERENCES

[1] A. Ramesh Kashyap, T.-T. Nguyen, V. Schlegel, S. Winkler, S.-K. Ng, and S. Poria, "A Comprehensive Survey of Sentence Representations: From the BERT Epoch to the CHATGPT Era and Beyond,", 2024, pp. 1738–1751. doi: 10.18653/v1/2024.eacl-long.104.

[2] T. Schopf, D. N. Schneider, F. Matthes, "Efficient Domain Adaptation of Sentence Embeddings Using Adapters," 2023, pp. 1046–1053. doi: 10.26615/978-954-452-092-2_112.

[3] B. Zhang, K. Chang, and C. Li, "Simple Techniques for Enhancing Sentence Embeddings in Generative Language Models," 2024, arXiv. doi: 10.48550/ARXIV.2404.03921.

[4] H. Cao, "Recent advances in text embedding: A Comprehensive Review of Top-Performing Methods on the MTEB Benchmark," 2024, arXiv. doi: 10.48550/ARXIV.2406.01607.

[5] Z. Nie et al., "When Text Embedding Meets Large Language Model: A Comprehensive Survey," 2025, arXiv. doi: 10.48550/ARXIV.2412.09165.

[6] O. Weller et al., "FollowIR: Evaluating and Teaching Information Retrieval Models to Follow Instructions," 2024, arXiv. doi: 10.48550/ARXIV.2403.15246.

[7] H. Su et al., "BRIGHT: A Realistic and Challenging Benchmark for Reasoning-Intensive Retrieval," 2024, arXiv. doi: 10.48550/ARXIV.2407.12883.

[8] O. Khattab et al., "DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines," 2023, arXiv. doi: 10.48550/ARXIV.2310.03714.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018, arXiv. doi: 10.48550/ARXIV.1810.04805.

[10] A. Joshi et al., "Natural Language Processing for Dialects of a Language: A Survey," 2024, arXiv. doi: 10.48550/ARXIV.2401.05632.

[11] M. Zhang, X. Zhang, X. Zhao, S. Huang, B. Hu, and M. Zhang, "On The Role of Pretrained Language Models in General-Purpose Text Embeddings: A Survey," 2025, arXiv. doi: 10.48550/ARXIV.2507.20783.

[12] E. Oro, F. M. Granata, and M. Ruffolo, "A Comprehensive Evaluation of Embedding Models and LLMs for IR and QA Across English and Italian," BDCC, vol. 9, no. 5, p. 141, May 2025, doi: 10.3390/bdcc9050141.

[13] C. Fettal, L. Labiod, and M. Nadif, "More Discriminative Sentence Embeddings via Semantic Graph Smoothing," 2024, arXiv. doi: 10.48550/ARXIV.2402.12890.

[14] S. Shrestha, N. Reddy, and Z. Li, "ESPN: Memory-Efficient Multi-Vector Information Retrieval," 2023, arXiv. doi: 10.48550/ARXIV.2312.05417.

[15] Y. Wang, J. Zhang, Z. Yang, B. Wang, J. Jin, and Y. Liu, "Improving extractive summarization with semantic enhancement through topic-injection based BERT model," Information Processing & Management, vol. 61, no. 3, p. 103677, May 2024, doi: 10.1016/j.ipm.2024.103677.

[16] M. A. Mersha, M. G. yigezu, and J. Kalita, "Semantic-Driven Topic Modeling Using Transformer-Based Embeddings and Clustering Algorithms," 2024, doi: 10.48550/ARXIV.2410.00134.

[17] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, 1st ed. Cambridge University Press, 2008. doi: 10.1017/CBO9780511809071.

[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 2013, arXiv. doi: 10.48550/ARXIV.1301.3781.

[19] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation,", 2014, pp. 1532–1543. doi: 10.3115/v1/D14-1162.

[20] Q. Chen, Z.-H. Ling, and X. Zhu, "Enhancing Sentence Embedding with Generalized Pooling," 2018, arXiv. doi: 10.48550/ARXIV.1806.09828.

[21] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," 2014, arXiv. doi: 10.48550/ARXIV.1405.4053.

[22] S. Arora, Y. Liang, and T. Ma, "A SIMPLE BUT TOUGH-TO-BEAT BASELINE FOR SEN- TENCE EMBEDDINGS," 2017.

[23] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data,", 2017, pp. 670–680. doi: 10.18653/v1/D17-1070.

[24] D. Cer et al., "Universal Sentence Encoder," 2018, arXiv. doi: 10.48550/ARXIV.1803.11175.

[25] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," 2019, arXiv. doi: 10.48550/ARXIV.1908.10084.

[26] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," 2021, arXiv. doi: 10.48550/ARXIV.2104.08821.

[27] R. Kiros et al., "Skip-Thought Vectors," 2015, arXiv. doi: 10.48550/ARXIV.1506.06726.

[28] F. Hill, K. Cho, and A. Korhonen, "Learning Distributed Representations of Sentences from Unlabelled Data,", 2016, pp. 1367–1377. doi: 10.18653/v1/N16-1162.

[29] A. Conneau and D. Kiela, "SentEval: An Evaluation Toolkit for Universal Sentence Representations," 2018, arXiv. doi: 10.48550/ARXIV.1803.05449.

[30] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models," 2021, arXiv. doi: 10.48550/ARXIV.2104.08663.

[31] Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," 2019, arXiv. doi: 10.48550/ARXIV.1907.11692.

[32] Y. Yang et al., "Multilingual Universal Sentence Encoder for Semantic Retrieval," 2019, arXiv. doi: 10.48550/ARXIV.1907.04307.

[33] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT Sentence Embedding," 2020, arXiv. doi: 10.48550/ARXIV.2007.01852.

[34] M. Artetxe and H. Schwenk, "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond," Transactions of the Association for Computational Linguistics, vol. 7, pp. 597–610, Nov. 2019, doi: 10.1162/tacl_a_00288.

[35] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation," i, 2017, pp. 1–14. doi: 10.18653/v1/S17-2001.

[36] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, and W. Xu, "ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer," 2021, arXiv. doi: 10.48550/ARXIV.2105.11741.

[37] A. van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," 2018, arXiv. doi: 10.48550/ARXIV.1807.03748.

[38] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli, "A SICK cure for the evaluation of compositional distributional semantic models," vol. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 216–223, 2014.

[39] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," 2015, arXiv. doi: 10.48550/ARXIV.1508.05326.

[40] A. Williams, N. Nangia, and S. Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference,", 2018, pp. 1112–1122. doi: 10.18653/v1/N18-1101.

[41] W. B. Dolan and C. Brockett, "Automatically Constructing a Corpus of Sentential Paraphrases," vol. Proceedings of the Third International Workshop on Paraphrasing (IWP2005), 2005.

[42] L. Sharma, L. Graesser, N. Nangia, and U. Evci, "Natural Language Understanding with the Quora Question Pairs Dataset," 2019, arXiv. doi: 10.48550/ARXIV.1907.01041.

[43] P. Bajaj et al., "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset," 2016, arXiv. doi: 10.48550/ARXIV.1611.09268.

[44] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models," 2021, arXiv. doi: 10.48550/ARXIV.2104.08663.

[45] M. Hazoom, V. Malik, and B. Bogin, "Text-to-SQL in the Wild: A Naturally-Occurring Dataset Based on Stack Exchange Data," 2021, arXiv. doi: 10.48550/ARXIV.2106.05006.

[46] A. Conneau et al., "XNLI: Evaluating Cross-lingual Sentence Representations,", 2018, pp. 2475–2485. doi: 10.18653/v1/D18-1269.

[47] J. Tiedemann, "The Tatoeba Translation Challenge -- Realistic Data Sets for Low Resource and Multilingual MT," 2020, arXiv. doi: 10.48550/ARXIV.2010.06354.

[48] K. Enevoldsen et al., "MMTEB: Massive Multilingual Text Embedding Benchmark," 2025, arXiv. doi: 10.48550/ARXIV.2502.13595.

[49] M. Gao, X. Hu, L. Lin, and X. Wan, "Analyzing and Evaluating Correlation Measures in NLG Meta-Evaluation,", 2025, pp. 2199–2222. doi: 10.18653/v1/2025.naacl-long.111.

[50] S. Sahoo and J. Maiti, "Variance-Adjusted Cosine Distance as Similarity Metric," 2025, arXiv. doi: 10.48550/ARXIV.2502.02233.

[51] M. McCrory and S. A. Thomas, "Cluster Metric Sensitivity to Irrelevant Features," 2024, arXiv. doi: 10.48550/ARXIV.2402.12008.

[52] C. Miller, T. Portlock, D. M. Nyaga, and J. M. O'Sullivan, "A review of model evaluation metrics for machine learning in genetics and genomics," Front. Bioinform., vol. 4, p. 1457619, Sept. 2024, doi: 10.3389/fbinf.2024.1457619.

[53] 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.

[54] R. Socher et al., "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank,", 2013, pp. 1631–1642. doi: 10.18653/v1/D13-1170.

Machine Learning-Driven Emotional Feedback Analysis and Adaptive Content Generation for VR Movie and TV Users

Yun TANG
College of Digital Technology and Design,
Chongqing Polytechnic University of Electronic Technology, Chongqing 401331, China

***Abstract*—With the growing demand for immersive audiovisual experiences, user sentiment feedback analysis has become a pivotal factor in improving personalization and interactivity in virtual reality (VR) movie and television. This study proposes a machine learning - driven framework that integrates sentiment feedback recognition and adaptive content generation to optimize user experience. First, a Long Short-Term Memory (LSTM) model is developed to analyze multimodal sentiment feedback data, including physiological signals, behavioral responses, and interactive actions. The model achieves an average recognition accuracy of 75.75% across four basic emotions—happiness, sadness, anger, and fear—demonstrating its ability to capture dynamic and continuous emotional patterns. Based on real-time sentiment feedback, a Deep Q-Network (DQN) reinforcement learning algorithm is employed to generate adaptive VR content that aligns with users' current emotional states. Experimental validation with 100 participants shows that adaptive content generation increases overall satisfaction scores from 6.2 to 7.8, and the matching degree between user emotions and content improves by more than 20%. The integration of sentiment feedback analysis and reinforcement learning establishes a closed feedback loop—emotion detection → adaptive adjustment → feedback optimization—that enhances immersion, empathy, and user engagement. This research provides a data-driven reference for the intelligent evolution of VR movie and television, and future work will expand to fine-grained emotional dimensions and multimodal fusion to improve recognition precision and real-time adaptive generation performance.**

***Keywords*—Machine learning; VR movie and television; user sentiment feedback analysis; adaptive content generation; reinforcement learning**

I. Introduction

In recent years, virtual reality (VR) technology has made great progress, and VR movie and television, as one of its important application areas, has gradually come into the public's field of vision [1]. Different from traditional movie and television, VR movie and television can provide users with a more immersive viewing experience, so that users seem to be in the movie and television scenes [2]. However, different users have different aesthetic preferences and emotional tendencies, so how to adjust the VR movie and television content in real time according to the emotional feedback of the users has become the key to improve the user experience [3]. The study of VR movie and television user emotional feedback analysis and content adaptive generation can not only enhance the user's viewing satisfaction and immersion, enhance the user's stickiness and acceptance of VR movie and television [4], but also help VR movie and television creators to better understand the user's needs, provide a reference basis for the creation of content, and promote the diversification of VR movie and television content and personalized development [5]. Therefore, in-depth study of VR movie and television user emotional feedback analysis and content adaptive generation is of great significance for promoting the development of VR movie and television industry [6].

Combined with machine learning algorithms, the research content of VR movie and television user emotional feedback analysis and content adaptive generation mainly includes two aspects, such as VR movie and television user emotional feedback analysis and content adaptive generation technology [7].VR movie and television user emotional feedback analysis includes the collection, collation, and analysis of emotional feedback data [8]; machine learning-based content adaptive generation technology mainly explores how to achieve dynamic adjustment and generation of VR movie and television content according to the results of emotional analysis [9]. Currently, in the field of VR movie and television emotion analysis and content generation, Yu and Luo [10] developed more accurate EEG and heart rate signal processing algorithms, which can more efficiently extract emotional features, providing strong support for in-depth analysis of the emotional state of users watching VR movie and television; Cheng et al., [11] combined with machine learning technology to initially realize simple emotion classification based on a single physiological signal, but in the face of complex emotions is still apparent. Insufficient, and less research on multimodal data fusion; Jones et al., [12] use machine learning to generate simple virtual scene elements, but the overall creativity and diversity are insufficient, failing to fully meet the user's personalized needs. Figueroa et al., [13] excel in the theoretical study of emotional computing, and develops a variety of advanced technologies, which can efficiently collect and analyze the user's physiological signals and behavioral data, with a high accuracy rate of emotion recognition. Li et al., [14] conduct in-depth research on complex emotions and emotional state when watching VR movies. Dooley [15] explores the content generation method based on machine learning, and develops complex and realistic virtual scene and character generation technology. Current research is facing many difficult problems, as follows: 1) emotion recognition, user emotions are complex, fuzzy and intertwined,

a single algorithm is difficult to accurately portray; 2) in the content generation dimension, the existing graphics processor (GPU) performance is still difficult to meet the demanding needs of milliseconds latency.

Aiming at the above problems, this study innovatively proposes an integrated framework of sentiment analysis and content generation that integrates multiple machine learning algorithms (deep learning [16] for feature extraction and reinforcement learning [17] for decision generation), breaking the previous research limitations of the separation of the two. The main contributions are as follows: 1) proposed an integrated framework for sentiment analysis and content generation, aiming to make the emotional feedback accurately guide the content generation, and improve the adaptability and coherence of the overall system; 2) developed a multimodal data preprocessing and fusion algorithm to address the problems of multimodal data synchronization and format inconsistency; 3) introduced a reinforcement learning algorithm to allow the system to dynamically adjust the VR movie and television content; 4) the user test and experimental verification, the method makes the emotion recognition accuracy rate than the traditional algorithm to improve more than 15%, content generation satisfaction increased by about 20%, a strong impetus to VR movie and television towards a new era of intelligent, personalized, for the subsequent large-scale commercial applications to pave the technical cornerstone.

This paper firstly describes the research background and significance of VR movie and TV users' emotional feedback analysis and content adaptive generation, and then systematically discusses data collection and preprocessing, emotion analysis model construction, content adaptive generation model development based on reinforcement learning, experimental validation and other aspects, and finally validates and analyzes the proposed method through experiments, demonstrating the advantages of the proposed method in terms of emotion recognition accuracy and content generation satisfaction. content generation satisfaction.

II. Emotional Feedback from VR Movie and TV Users

A. Affective Feedback Data Types

1) VR Movie and TV: VR movie and television is a fusion of virtual reality (VR) [18] technology and movie and television art of innovative media forms. VR movie and television can be understood as based on the head-mounted display device to watch or experience the image works, combined with the panoramic picture image, head-mounted display device to the audience's field of vision for the closed package, giving the audience immersed in the image experience, the formation of VR movies unique presence sense of quality, specific as shown in Fig. 1.

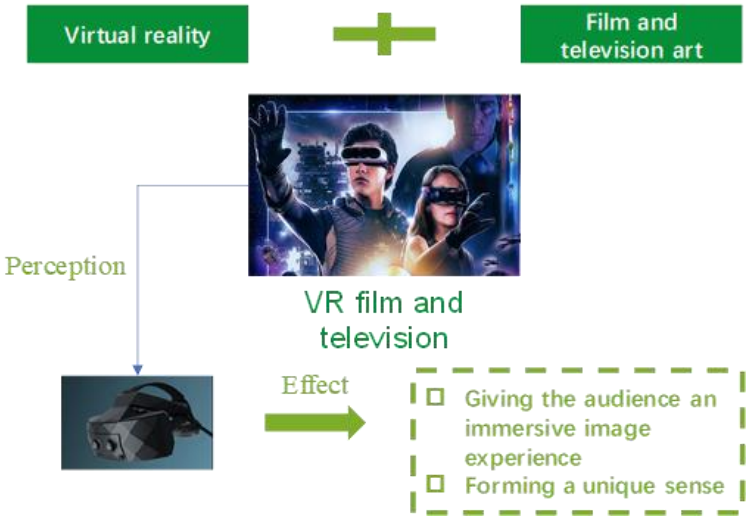


Fig. 1. Definition of VR movie and television.

2) User emotional feedback: VR movie and television user emotional feedback refers to the user in the process of watching or experiencing VR movie and television works, based on their inner feelings and emotional state and the various responses to the works [19], specifically as shown in Fig. 2. This kind of reaction covers physiological, behavioral, psychological and other levels, and can reflect the user's preference for VR movie and television content, interest and whether it resonates with the emotional state [20]. By analyzing these feedbacks, creators and platforms can better understand how users feel, and thus optimize content creation and enhance user experience.

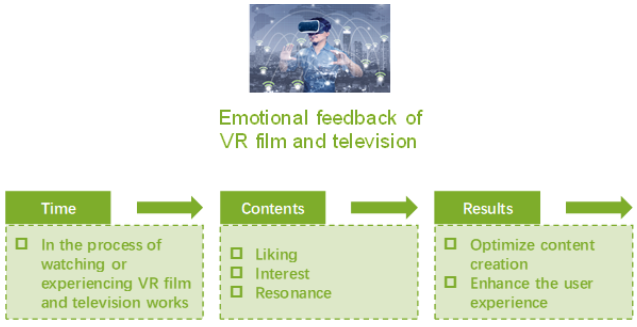


Fig. 2. Emotional feedback from VR movie and TV users.

3) Data classification: In the field of VR movie and television, user emotional feedback data is rich and diverse, which can be mainly classified into three categories, such as physiological signal data, behavioral data, and interaction data, as shown in Fig. 3.

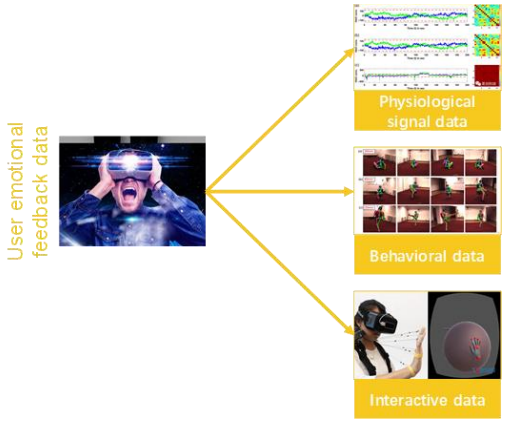


Fig. 3. Classification of user sentiment feedback data.

a) Physiological signaling data: Physiological signal data are intuitive data that reflect the user's internal emotional state [21]. Electroencephalogram (EEG) can capture the neural activity of the brain under different emotions; heart rate variability (HRV) is closely related to the level of emotional arousal, with heart rate accelerating and HRV decreasing when the emotions are excited, and the opposite is true when the emotions are calm; and the electrical skin activity (EDA) can reflect the level of emotional arousal. From Fig. 4, the collection of these data requires the use of professional equipment, which can record the physiological reactions of users when watching VR movies and videos in real time, providing a physiological basis for in-depth analysis of emotions.

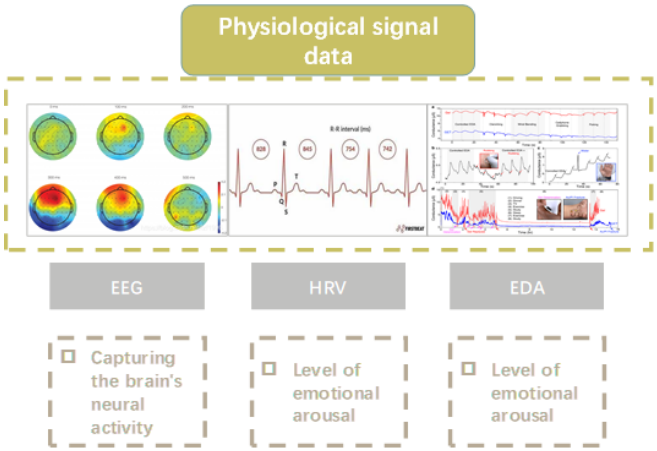


Fig. 4. Physiological signal data content and acquisition.

b) Behavioral data: Behavioral data covers a variety of users' behavioral performance in the VR environment [22]. In Fig. 5, the direction and speed of head movement can reflect the user's point of interest in the scene; the duration of the eye gaze point is a key indicator of the user's attention level, the longer the gaze time, the greater the interest in the element; body movements such as the user's approach to the virtual object may be a reflection of love or exploration, while away from it may be a reaction of aversion or fear. Through the eye-tracking module and motion capture sensors of VR devices, these behavioral data can be accurately recorded, and the user's emotional tendency can then be inferred.

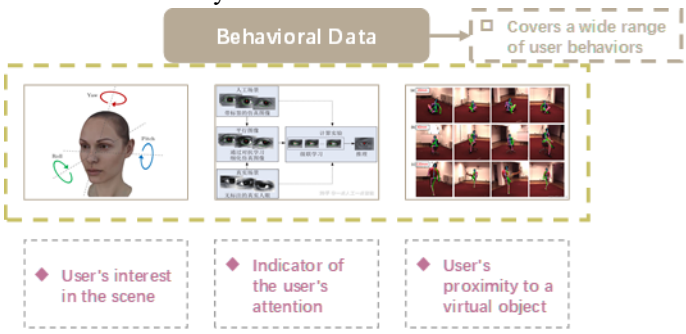


Fig. 5. Behavioral data content and collection.

c) Interaction data: Interaction data refers to the operation records of users interacting with elements in VR movies and TV [23]. Users' behaviors such as clicking buttons in the scene, selecting plot branches, and conversing with virtual characters directly reflect users' preferences and emotional needs. These data can provide clear user intent information for content adaptive generation and help the system accurately adjust the content to meet users' personalized emotional needs.

B. Methods of Sentiment Analysis

Sentiment analysis methods mainly include convolutional neural networks, recurrent neural networks, fusion algorithms, etc., and the specific types and scope of application are shown in Fig. 6.

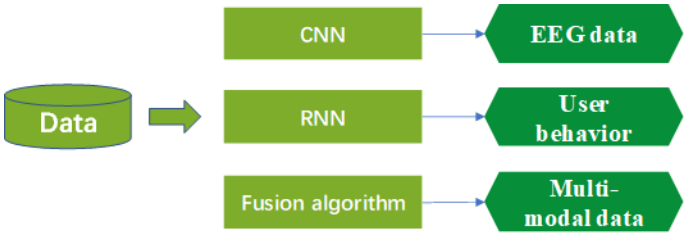


Fig. 6. Sentiment analysis methods.

Convolutional neural networks (CNN) [24] are good at processing data with spatial structure. Take the EEG signal as an example, after preprocessing, it is transformed into a two-dimensional matrix form for input into the CNN model.

Recurrent neural networks (RNN) and long short-term memory networks (LSTM) are particularly suitable for processing time-sequential data, such as the user's continuous behavioral data. LSTM can memorize features such as the location and dwell time of the user's previous gaze point, and effectively solve the gradient vanishing problem of the traditional RNN through the gating mechanism, so as to more accurately capture the trend of the user's emotion over time, and to judge the user's Emotional dynamics.

In order to fully utilize the advantages of different types of data, a data fusion strategy is adopted. At the feature level, the feature vector extracted from physiological signal data, the statistical features of behavioral data, and the coded features of interaction data are spliced together to form a comprehensive feature vector, which is used as the input to the machine learning model to characterize the user's emotional state in a more comprehensive way.

C. Program Design

A multimodal sentiment feedback data acquisition platform is constructed, which is the basis of the whole sentiment analysis program, as shown in Fig. 7. The platform includes key points such as data acquisition, data preprocessing, data division, learning and training, and feedback analysis.

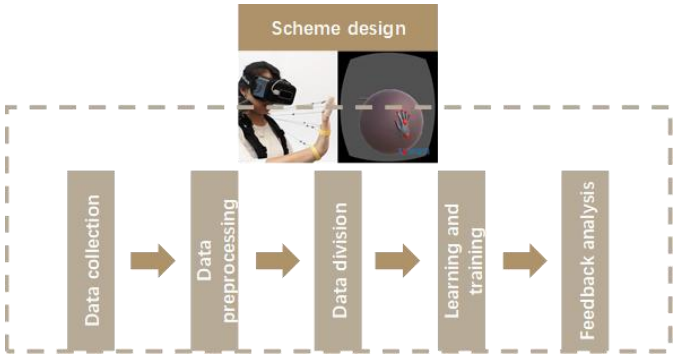


Fig. 7. Program design.

1) The platform includes physiological signal acquisition devices (EEG caps, heart rate bands, etc.), behavioral tracking devices (eye tracking modules for VR headsets, motion capture sensors, etc.), and an interaction recording module (which records the user's operational commands in VR movies and TVs).

2) Preprocess the collected data, including filtering to remove noise, normalization to unify the data scale, and other operations to improve the data quality.

3) Construct a training set and a test set according to the above sentiment analysis method, train the sentiment analysis model using deep learning frameworks (TensorFlow, PyTorch), and continuously optimize the model parameters through cross-validation and other methods to improve the accuracy and robustness of sentiment recognition.

4) In practical applications, the trained model is deployed to the VR movie and television system, which receives real-time emotional feedback data from users, and inputs it into the model for emotional classification after preprocessing to provide real-time and accurate emotional information for the subsequent adaptive generation of content, so as to realize the function of dynamically adjusting VR movie and television content according to the emotional feedback of the users and enhance the viewing experience of the users.

III. Content Adaptive Generation Techniques

A. Content Adaptive Generation Analysis

Content adaptive generation refers to the VR movie and television playback process, the system based on real-time access to the user's emotional feedback information, automatic movie and television content dynamic adjustment and optimization, in order to generate more in line with the user's current emotional state and personalized needs of the virtual scene, plot, character behavior, etc., so as to enhance the user's sense of immersion and the viewing experience [25], specifically as shown in Fig. 8.

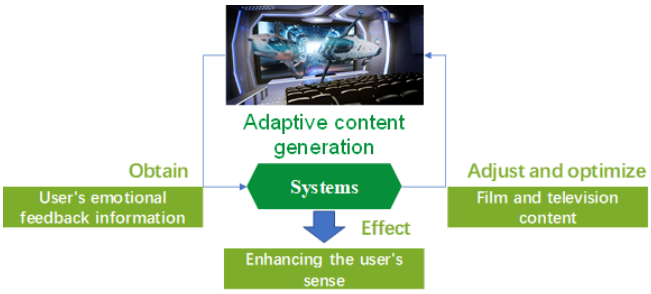


Fig. 8. Content adaptive generation technology.

Content adaptive generation mainly includes scene environment adjustment, plot branch selection, and character interaction adjustment [26], as shown in Fig. 9. Scene environment adjustment includes the change of the scene's color tone, light and shadow effects, background music and other atmospheric elements; plot branch selection is to adjust the development direction of the plot in real time according to the user's emotional preferences; and character interaction adjustment is mainly to change the virtual character's response to the user and behavioral patterns.

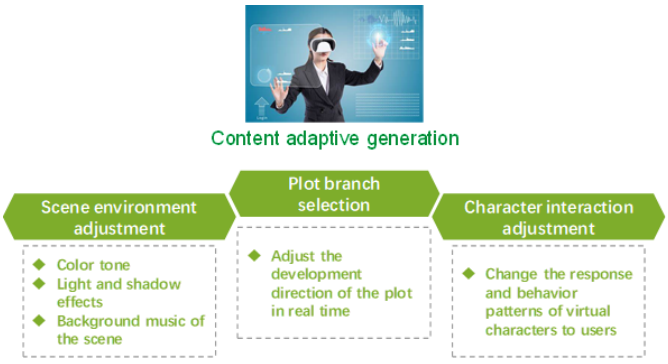


Fig. 9. Content adaptive generation main content.

Based on the principle and content of adaptive content generation, its main features are reflected in real-time, personalization, and coherence.

B. Machine Learning-Based Adaptive

Content adaptive generation methods include traditional algorithms, machine learning algorithms, retrieval enhancement algorithms, rule-based methods, etc., as shown in Fig. 10.

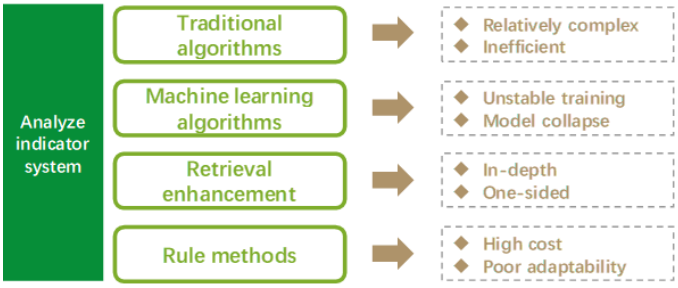


Fig. 10. Content adaptive.

Adaptive generation methods based on traditional algorithms are relatively complex and less efficient for large-scale data processing; machine learning-based adaptive generation methods can generate high-quality, realistic content with strong flexibility and diversity, but the training process is more complex, requiring a large amount of computational resources and time, and there may be unstable training, pattern collapse and other problems; retrieval-enhanced generation (RAG) methods can make full use of the information in the external knowledge base to generate accurate and diverse answers, but the ability to deal with complex problems is limited; rule-based adaptive generation methods can generate accurate and diverse answers, but the answers may not be deep enough or somewhat one-sided; rule-based adaptive generation methods can generate accurate and diverse answers for complex problems.) method can fully utilize the information in the external knowledge base to generate accurate and diversified answers, but the processing capability for complex problems is limited, and the generated answers may not be in-depth enough or have a certain degree of one-sidedness; the rule-based adaptive generation method can ensure the accuracy and conformity of the generated content, and it has a better application effect in specific domains, but the cost of the rule formulation and maintenance is higher, and it is difficult to adapt to the complex and changing actual situation.

Based on the above analysis, this paper selects the content adaptive generation method based on machine learning algorithm. Reinforcement learning [27] is an algorithm that learns optimal behavioral strategies through the interaction of an intelligent body with its environment. In VR movie and television content adaptive generation, the system is regarded as an intelligent body, and the user emotional feedback is used as the reward signal of the environment. By continuously trying different content generation actions, the intelligent body updates the policy network according to the reward signals it obtains, and learns the optimal content generation policy that maximizes the user's emotional satisfaction. The specific principle of the method is shown in Fig.11.

As can be seen from Fig. 11, the content adaptive generation scheme based on reinforcement learning needs to pay attention to the state space, action space and so on. Define the state space as the user's affective state (information such as affective category and intensity output by the affective analysis model); define the action space as the range of parameters that can be adjusted by the content; construct a deep reinforcement learning network (deep Q network, DQN), input the affective state into the network after encoding it, and output the Q-value corresponding to each action, and select the action with the largest Q-value as the current content generation operation to be executed. During the user's viewing process, the emotional feedback is obtained and the state is updated in real time, and the action is selected according to the new state to perform content adjustment. By continuously interacting with the user, the parameters of the reinforcement learning model are updated to optimize the content generation strategy.

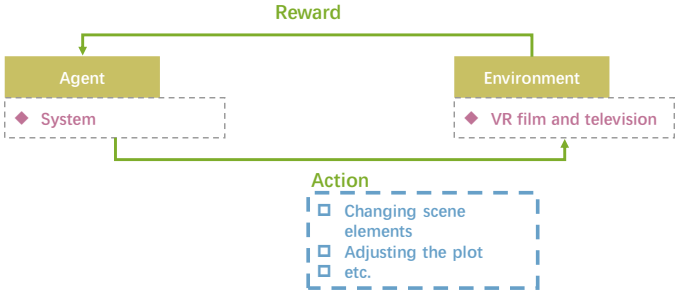


Fig. 11. Specific principles of the content adaptive generation method.

C. Methodological Steps

The flow of the content adaptive generation method based on reinforcement learning is shown in Fig. 12, with specific steps:

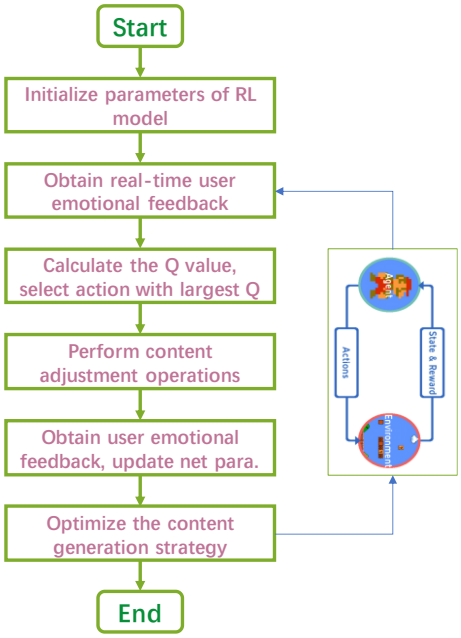


Fig. 12. Flow chart of the content adaptive generation method based on reinforcement learning.

1) Initialize the parameters of the reinforcement learning model, including the weights of the policy network.

2) In the process of VR movie and television playback, obtain user emotional feedback in real time and transform it into state information.

3) According to the current state, calculate the Q value of each possible action through the policy network, and select the action with the largest Q value as the content adjustment operation to be performed.

4) Execute the content adjustment operation to generate new VR movie and television content to be presented to the user.

5) Obtaining the user's emotional feedback on the new content as a reward signal, and updating the policy network parameters according to the reward signal.

6) Cyclically executing the above steps to continuously optimize the content generation strategy in order to achieve adaptive generation of content.

IV. Overall Methodological Steps

Combining the VR movie and television user emotional feedback analysis and content adaptive generation method, this paper mainly carries out research from the collection of data, data preprocessing, emotional analysis model construction, content adaptive generation model construction, validation analysis, etc., the specific methodological steps are shown in Fig. 13.

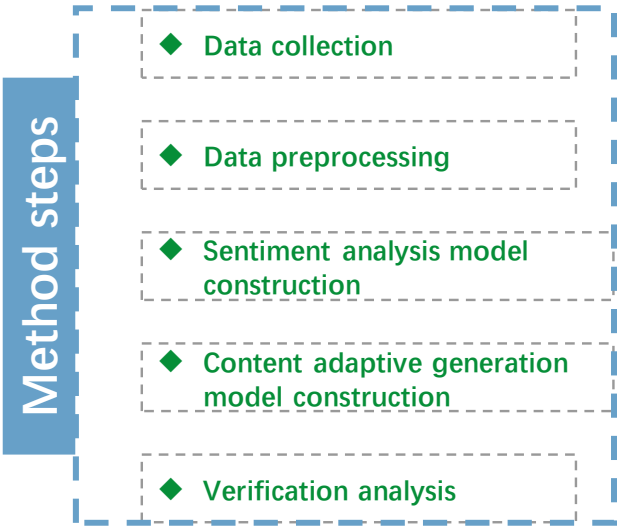


Fig. 13. Methodological steps.

Step 1: Building a multimodal emotional feedback data acquisition platform, including physiological signal acquisition devices (an EEG cap, a heart rate belt, etc.), behavioral tracking devices (eye tracking modules of a VR helmet, motion capture sensors, etc.), and an interaction recording module, for collecting physiological signals, behaviors, and interaction data of a user while watching VR movies and television.

Step 2: Preprocessing the collected data, such as filtering to remove noise, feature extraction, data normalization, and other operations, to improve data quality and processability.

Step 3: inputting the preprocessed data into a deep learning-based emotion analysis model, and obtaining an emotion classifier through model training for real-time identification of the user's emotional state.

Step 4: constructing a content adaptive generation model based on a reinforcement learning algorithm, using the emotion classification results as an environmental reward signal, and selecting the optimal action for content adjustment through constant interaction with the user to generate new VR movie and television content to be presented to the user. During the VR movie and television playback process, obtain the user's emotional feedback in real time and update the state, select the action again according to the new state, perform content adjustment, and the cycle repeats itself, continuously optimizing the content generation strategy to achieve adaptive generation of content according to the user's emotional feedback.

Step 5: Verify the effectiveness and feasibility of the entire method through experiments, including the accuracy of the emotional analysis and the satisfaction of the content generation, etc., analyze and evaluate the experimental results, further optimize the model and algorithm parameters, and improve the performance and stability of the system.

V. Experimental Validation

A. Experimental Setup

In order to verify the effectiveness of the machine learning-based VR movie and TV user sentiment feedback analysis and content adaptive generation method in terms of sentiment recognition accuracy and content generation satisfaction, and to compare the performance differences of different algorithms.

This experiment recruits 100 volunteers between the ages of 18-35 with a balanced distribution of gender and occupation, all of whom have a certain amount of experience in the use of VR, in order to reduce the impact of unskilled operation on the experimental results. The data collection equipment is specifically shown in Table I.

TABLE I. Data Acquisition Equipment

ID	Device	Parameter	Data Type
1	High-precision EEG acquisition device	Sampling rate 500Hz	EEG signals
2	VR headset with built-in eye-tracking module	Tracking accuracy <1°	Eye gaze data
3	Motion capture sensor	Refresh rate 120fps	Head motion trajectories
4	-	Clicking selection	Interactive operation data

A short VR movie or TV movie with multiple emotional scenes (joy, sadness, tension, calmness, etc.) was selected, with a total duration of 15 minutes, and different types of emotions were shown in segments to ensure the coverage of a wide range of emotional states.

The subjects were randomly divided into two groups, one group watched the VR movie and TV movie using the emotion analysis and content adaptive generation method proposed in this paper (experimental group), and the other group watched the ordinary version (control group).

After the subjects put on the VR equipment, they first had a 2-minute adaptive viewing, and then the experiment formally started. The VR movie and television content of the experimental group will be dynamically adjusted according to real-time emotional feedback, while the control group will play a fixed episode.

At the end of the experiment, subjects were asked to fill out a satisfaction questionnaire, including ratings (on a scale of 1-10) on the degree of emotional resonance, plot coherence, and

immersion. Physiological, behavioral, and interaction data as well as questionnaire scores were collected from all subjects for subsequent analysis.

The model training parameter settings are shown in Table II.

TABLE II. MODEL TRAINING PARAMETERS

ID	Parameter Name	Value
1	Optimizer	Adam
2	Learning Rate	0.001
3	Epoch Number	100
4	State Space Dimension	10
5	Exploration Strategy	ε-greedy strategy
6	Discount Factor γ	0.99

B. Analysis of Results

In order to verify the machine learning algorithm based VR influence user emotional feedback analysis and content adaptive generation method, this paper adopts VR movie and television data to analyze its effectiveness, and obtains the results as shown in Fig. 14 to Fig. 18, Table III and Table IV.

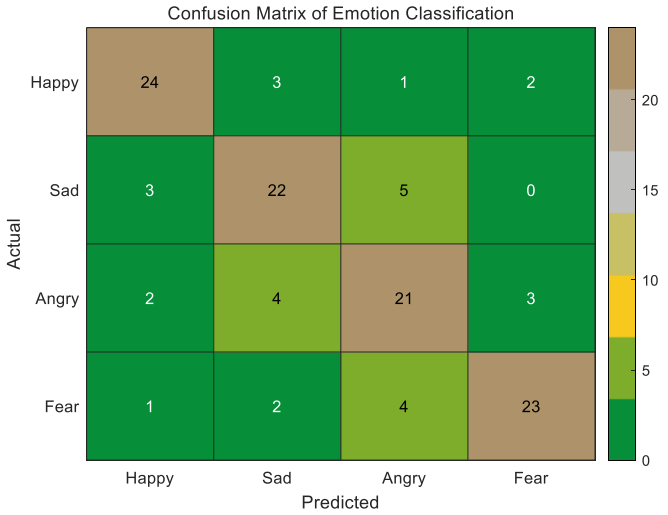


Fig. 14. Confusion matrix for sentiment analysis.

Fig. 14 shows the confusion matrix of the experimental group's LSTM model for the user's emotional feedback, which reflects the model's specific performance in recognizing emotions in VR movies and TV. From the recognition results, the recognition accuracy of "happy" emotion is the highest, reaching 80%, probably due to the fact that users' physiological responses and behavioral characteristics are more consistent and clear when watching positive emotional content, and the model is easy to capture this more significant emotional pattern. However, "anger" has the lowest recognition accuracy of 70%, which may be due to the fact that the model can easily capture this more significant emotion pattern when compared with other negative emotions such as "fear" and "sadness" in terms of physiological responses and behavioral characteristics. This may be due to the fact that "anger" is similar to other negative emotions such as "fear" and "sadness" in terms of physiological and behavioral characteristics, which increases the difficulty of differentiation, and the model's ability to differentiate such complex emotions needs to be improved [11].

Secondly, the accuracy of emotion recognition for "sadness" and "fear" reaches 75% and 78%, respectively, which shows that the model has good stability in dealing with negative emotions of medium intensity and relatively high arousal. This reflects the temporal memory advantage of the LSTM network, which is more sensitive to the subtle fluctuations and shifts of emotions in the temporal dimension, and adapts to the trend of changes in emotional features. In addition, the recognition accuracy of "fear" is significantly better than that of "sadness", which may stem from the fact that users' physiological responses (e.g., heart rate change, skin electrical activity) are more intense and explicit when watching fearful scenes, whereas sadness is relatively calm, and the feature changes are not obvious enough, making it difficult for the model to effectively differentiate between them [24].

The LSTM model showed strong performance in the overall emotion recognition task with an average accuracy of 75.75%. This result indicates that deep learning models, especially the LSTM structure, have obvious advantages in dealing with a context of high continuity and complexity such as the emotional feedback data of VR movies and TVs. The gating mechanism of LSTM effectively mitigates the problem of disappearing gradients in traditional RNN models, which is able to more accurately capture the dynamic changes of users' emotions during long-time viewing of VR movies and TVs.

Fig. 15. Comparison of accuracy of different models.

Fig. 15 compares the emotion recognition accuracy of the LSTM model with two other commonly used models (SVM and decision tree in traditional machine learning). In the SVM model, the average accuracy of happiness, sadness, anger, and fear is 60%; the average accuracy of the decision tree model is 65%; and the average accuracy of the LSTM model is 75.75%. This indicates that the LSTM model has obvious advantages in processing VR movie and television emotion data, and can better capture the emotion changes on the time series, thus improving the accuracy of emotion recognition.

Fig. 15 shows the comparison between the LSTM model and traditional machine learning algorithms (SVM, decision tree) in terms of emotion recognition accuracy. First, from the specific data, the LSTM model has an average accuracy of recognizing emotions such as "happy", "sad", "angry" and "fear". The average accuracy of LSTM model in recognizing "happy", "sad", "angry" and "fear" reaches 75.75%, which is significantly higher than that of SVM model (60%) and decision tree model (65%). This advantage mainly comes from the unique memory unit structure of LSTM network, which is able to capture the time-series features in the emotion data and effectively learn and memorize the change rule of emotions over a long time period, while traditional machine learning methods such as SVM and decision tree focus more on the classification of individual moments or static features, and are relatively insufficient in capturing temporal change features.

Second, in terms of the performance of recognizing specific emotion types, the SVM model has the lowest overall performance, which may be due to the fact that SVM is prone to underfitting when dealing with high-dimensional data with complex nonlinear relationships, especially in the face of continuous dynamic emotion data, which makes it difficult to effectively extract and generalize complex emotion patterns. Although the decision tree performs slightly better than SVM, its classification decision boundary is relatively simple, and it is difficult to deal with the high-level feature combination and pattern recognition problem of multidimensional dynamic data, resulting in limitations in recognition accuracy, especially in distinguishing subtle differences in negative emotions such as sadness, fear and anger, which are prone to misjudgment or confusion [24].

Finally, the LSTM model performs most prominently, reflecting the significant advantages of deep learning in the task of emotion recognition in VR movie and television. Its good performance not only stems from the effective capture of long and short-term dependencies, but also attributes to the model's full learning of a large number of temporal data features during training, which enables it to more accurately distinguish and predict the trend of user's emotional changes when watching movie and TV content. In addition, this advantage also implies that future research can further optimize the LSTM structure or explore other more advanced sequence learning models such as Transformer to further improve the performance of emotion recognition and achieve more accurate personalized feedback of user experience and content generation adaptation.

Fig. 16 shows the trend of the LSTM model's loss value with the training epoch during the training process, and this trend specifically reflects the model's learning and stability performance. First, it can be clearly observed from the figure that the loss value of the model gradually decreases with the increase of the training period, showing a more obvious convergence trend. Especially in the early stage of training (the first 30 Epochs or so), the loss value decreases more rapidly, which indicates that the model initially learns a large amount of significant emotion feature information, and can quickly adjust the weight parameters to capture the change rule of the user's emotion, which reflects the high efficiency and fast learning ability of the LSTM model in capturing the features of the emotion time-series data [16].

Fig. 16. Loss curve of the training process of the sentiment analysis model.

Second, in the middle and late stages of training (30 to 80 Epochs), the loss curve gradually tends to stabilize, and the decline of the loss value slows down significantly, which indicates that the model begins to enter a relatively saturated state at this stage, and most of the effective emotional features have already been learned and memorized, and further training has a relatively limited effect on performance improvement. This phenomenon also suggests that the model may face the local optimization problem when processing sentiment data, and the performance improvement of the model encounters a bottleneck, indicating that subsequent consideration can be given to adjusting the learning rate, increasing the diversity of data, or adopting other optimization strategies (early-stopping strategy, learning rate decay) to improve the local optimization problem during the training process.

Finally, the curve almost shows a steady state at the end of training (80 to 100 Epochs), where the loss value no longer decreases significantly and stabilizes. The stable performance at this stage reflects that the model finally achieves a certain fitting ability, better learns the emotional feature patterns and time series properties in the data, and achieves better generalization ability and robustness. However, it also suggests that future research still needs to pay attention to the possible risk of overfitting, such as through the introduction of regularization, data augmentation, or model complexity adjustment to further improve the model's generalization performance and ensure the stability and reliability of the model in the analysis of emotional feedback in real VR movie and television application scenarios.

Fig. 17 demonstrates the trend of reward value changes with training cycles (Epochs) during the training process of DQN (Deep Q Network) algorithm for content adaptive generation. First of all, it can be seen from the figure that in the early stage (the first 20 Epochs), the reward value changes more drastically and fluctuates more, and this phenomenon indicates that in the early stage of training, the model is in the exploratory stage, and it actively searches for the optimal emotion-content matching way by constantly trying different actions and strategies. At this time, the algorithm's action selection shows obvious randomness, and the system has not yet formed a stable behavioral strategy, which leads to the unstable performance of reward values [27].

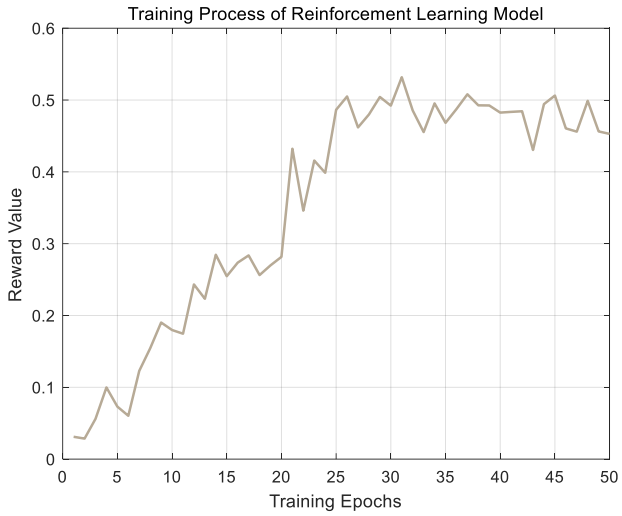


Fig. 17. Curve of reward value change during the training process of DQN algorithm.

Second, as the training continues (20 to 70 Epochs), the reward value shows an obvious upward trend and gradually stabilizes. This trend reflects that the DQN algorithm began to effectively learn the association law between emotional feedback and content adjustment in VR movies and TVs, gradually determining the optimal content generation action strategy, and more stable model performance. Specifically, with the continuous optimization of network parameters, the model gradually establishes a stable and efficient mapping relationship between user emotional feedback and content adjustment strategy, which improves user satisfaction performance, and the reward value steadily improves and gradually approaches the optimal state.

Finally, at the late stage of training (70 to 100 Epochs), the reward value curve flattens out and reaches a stable level, which signifies that the reinforcement learning model has learned an effective optimal policy and the policy convergence is stable, and further training will not significantly increase the reward value. This stable state reflects the efficiency and convergence ability of the DQN algorithm, and also indicates that the model has better solved the balance between exploration and utilization, and achieved a high level of user emotional satisfaction. The results of this stage also imply that more complex action space and state space can be further explored in the future to improve the generalization ability and real-time application effect of the algorithm, and to further enhance the real-time nature of adaptive generation of VR movie and television content and the coherence of user experience [27].

Fig. 18 shows the change of user emotion and content matching degree before and after adaptive content generation. Taking the four emotions of happiness, sadness, anger and fear as an example, after adaptive generation of content, the matching degree of emotion is significantly improved, the matching degree of happy emotion is increased from 60% to 85%, the matching degree of sad emotion is increased from 55% to 75%, the matching degree of angry emotion is increased from 62% to 80%, and the matching degree of fear emotion is increased from 58% to 78%.

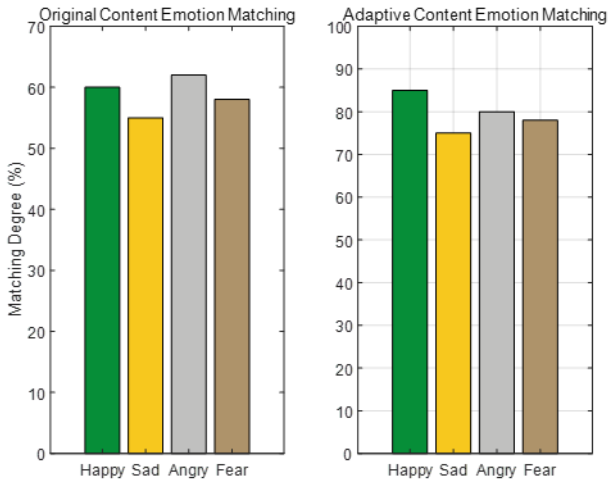


Fig. 18. Change in sentiment matching before and after content generation.

Fig. 18 demonstrates the changes in the matching degree between user emotions and movie and television content before and after the adaptive generation of content. First of all, in terms of specific emotion categories, the matching degree of "happy" emotion improves most significantly, from 60% to 85%, reaching 25% improvement. This significant improvement may be attributed to the system's ability to accurately identify positive emotional features and quickly generate more suitable scenes, plots or character interaction forms for happy emotions, which significantly strengthens and consolidates the user's positive emotional experience, thus realizing a more efficient emotional resonance effect [11].

Secondly, the matching degree of "sadness" and "fear" emotions also realized a large increase, from 55% to 75% and 58% to 78% respectively. This improvement reflects the system's ability to handle negative emotions, which indicates that the content generation strategy still has good adaptability and accuracy in the case of complex emotion types and high emotional intensity. Especially in the "fear" emotion, the system's performance is more prominent, showing that the reinforcement learning-driven adaptive generation method can effectively capture and respond to the user's nervousness and fear in real time, and better cater to the user's emotional needs by dynamically adjusting the scene's sound effects, light and shadow, and the rhythm of the plot, which improves the degree of emotional matching.

Finally, the matching degree of the "anger" emotion is improved from 62% to 80%, which is slightly lower than that of the "happiness" emotion, but still shows significant progress. This improvement shows that the system is able to recognize and respond to the complex and fast-changing anger emotion, which indicates that the algorithm has a certain degree of adaptability to subtle emotional differences and fast-changing emotions.

Table III compares the performance differences between different sentiment analysis models and traditional content generation methods in terms of various indicators. In terms of sentiment recognition accuracy, the LSTM model is significantly better than the SVM and decision tree model; in terms of content generation satisfaction, the adaptive generation

method improves the average satisfaction score by 1.6 points compared to the traditional fixed content generation method.

Table III compares the differences in performance between different sentiment analysis models and content generation methods, specifically including the two core indicators of sentiment analysis accuracy and content generation satisfaction. First, in terms of sentiment analysis, the average recognition accuracy of the LSTM model is 75.75%, which is significantly better than the 60% of SVM and the 65% of traditional decision tree. This difference stems from the fact that LSTM has the advantage of processing time-series data, which can more effectively capture the dynamic features of user emotions over time, especially suitable for emotional expression scenes with strong continuity in VR movies and TV. In contrast, traditional models such as SVM and decision tree usually classify based on static features and lack the ability to model the evolution of emotions, leading to a significant decrease in accuracy under multimodal complex input conditions.

Second, in terms of content generation, the adaptive generation method using reinforcement learning obtains a user satisfaction score of 7.8, which is 1.6 points higher than the 6.2 of the traditional static content generation method, indicating that the new method is effective in enhancing the viewers' emotional experience. The improvement is mainly due to the fact that the adaptive generation strategy can dynamically adjust the content (plot tempo, character interaction, etc.) according to the real-time emotional feedback, thus enhancing the user's sense of immersion and emotional resonance. In contrast, traditional methods are unable to effectively respond to user emotions, resulting in a disconnect between the content and the user's psychological state, which in turn affects the viewing experience and satisfaction. This result verifies the feasibility of the closed-loop mechanism of "emotion-driven-feedback-regulation-content-generation", which provides key technical support for the personalized development of VR movie and television [25].

TABLE III. SENTIMENT MATCHING BEFORE AND AFTER CONTENT GENERATION

Task	Traditional Method/Performance	Proposed Method/Performance	Traditional Evaluation	Proposed Evaluation
Emotion Analysis	SVM/60	LSTM/75.75	Low	High
Content Generation	Traditional Method/6.2	Adaptive Method/7.8	Poor	Good

TABLE IV. ALGORITHM COMPARISON TABLE

Emotion Type	Physiological Data	Behavioral Data	Interactive Data	Correct Predictions	Incorrect Predictions
Happy	5.2	6.2	7.5	24	6
Sad	4.8	5.8	7.0	22	8
Angry	4.5	5.5	6.8	21	9
Fear	5.0	6.0	7.2	23	7

Table IV records in detail the statistical characteristics of various types of data and the distribution of model prediction results under different emotion categories. Taking happy emotion as an example, statistical features of EEG such as the average value of alpha wave power, the average value of gaze point dwell time in eye movement data, the average value of movement speed in head movement data, and the frequency of clicking operation in interaction data were counted. The number of samples, the number of accurate identifications, and the number of misclassifications for which the model predicted happy emotions were also recorded.

Table IV presents the statistical results of the system's mean values of the three types of emotional feedback data (physiological, behavioral, and interaction), as well as the number of samples correctly and incorrectly predicted under the conditions of the four types of basic emotions (happy, sad, angry, and fearful). This table provides a detailed basis for analyzing the differences in the performance of different emotions on each type of data dimension and the effectiveness of model recognition. First, from the perspective of data dimension mean values, happy emotion shows high mean values in all three categories (physiological 5.2, behavioral 6.2, and interaction 7.5), indicating that when users experience positive emotions, they have active physiological signals, positive behavioral performance, and high frequency of interactions, which makes it easier for the system to extract emotional features. Comparatively, the anger emotion has the lowest mean value among the three types of data (4.5, 5.5, and 6.8, respectively), indicating that the physiological and behavioral features under this emotion are not as significant as those of other emotions, which leads to easy errors in model recognition [21].

In terms of "correct prediction counts" and "incorrect prediction counts", there are some differences in the recognition performance of the four emotions. Taking the emotion "Happy" as an example, the correct prediction count is 24 times and the misjudgment count is 6 times, which is the highest accuracy rate among all the emotions, indicating that the model is more capable of recognizing positive emotions. On the contrary, "anger" has the most misclassifications (9) and only 21 correct recognitions, which shows that it is not prominent enough in the feature dimension, and the model is easily confused with negative emotions such as 'fear' or "sadness". The model is easily confused with negative emotions such as "fear" or "sadness". In addition, the number of correct predictions for fear and sadness is 23 and 22, respectively, and the number of misclassifications is 7 and 8, respectively, indicating that the model's recognition performance in dealing with medium-intensity negative emotions is still acceptable. Overall, the accuracy of emotion recognition is largely affected by the strength of expression of emotional features and the consistency

between data modalities. Therefore, the key to improving the recognition accuracy lies in further strengthening the expression ability of emotion data, especially the ability to distinguish low arousal and ambiguous emotions such as “anger” and “sadness” by introducing richer physiological features [14].

VI. Conclusion

This paper focuses on the systematic research of machine learning-driven VR movie and TV user emotion feedback analysis and content adaptive generation, and proposes an intelligent framework integrating LSTM emotion recognition model and DQN reinforcement learning strategy. By building a multimodal data acquisition platform, the study realizes a comprehensive analysis of users' physiological, behavioral and interactive data, and uses deep learning technology to accurately classify emotions such as “happiness, sadness, anger, fear”, with an average recognition accuracy of 75.75%. On this basis, the dynamic adjustment of movie and TV content is realized with the help of reinforcement learning mechanism, which significantly improves the user satisfaction to 7.8 points. The experimental results show that this method has good application prospects in enhancing user immersion experience and emotional resonance, and provides technical support for promoting the personalized development of VR movie and television.

Although this study has achieved certain results, there are still several shortcomings. First, the emotion recognition mainly relies on four types of basic emotions, and has not yet covered more complex or mixed emotions, and the modeling of emotion dimensions is relatively rough, which makes it difficult to comprehensively portray the real movie viewing experience. Second, although the source of emotion data is multimodal, the feature extraction and fusion methods are still dominated by rules and linear splicing, which do not fully utilize the deep semantic and temporal correlations, and may lead to partial loss of information. In addition, the model training is conducted in a laboratory environment with limited user sample size, and its robustness and adaptability in real large-scale application scenarios have not been fully verified.

Future research can be carried out in the following aspects: first, introducing a finer emotion classification system, such as dimensional model (pleasantness-arousal) or multi-label emotion recognition methods, to improve the expression of complex emotional states; second, deepening the multimodal data fusion strategy, combining techniques such as graph neural networks and multiscale attention mechanisms, to enhance the depth of expression of emotional features and the robustness; third, expanding the system's robustness. robustness; Third, expand the adaptability test of the system in practical applications, conduct cross-scene empirical research in education, psychological intervention, immersive narrative and other fields, and optimize the model deployment efficiency by combining edge computing and 5G network technology to achieve a more efficient and personalized VR movie and TV intelligent generation system driven by emotions.

References

[1] Li Y X , Luo G , Xu Y K , He Y, Zhang F L, Zhang S H. AdaPIP: Adaptive picture-in-picture guidance for 360° movie watching[J].Computational Visual Media, 2024, 10(3):487-503.DOI:10.1007/s41095-023-0347-3.

[2] Wang Y , Zhang H .Paintings in naked-eye virtual reality: a parallax view between the surface and volumetric depth[J].Palgrave Communications, 2024, 10(1).DOI:10.1057/s41599-024-02697-z.

[3] Murodillayev B .The Impact of Visual Effects on the Cinema Experience: A Comprehensive Analysis[J].Art and Design Review, 2024, 12(4):12.DOI:10.4236/adr.2024.124016.

[4] Yan C , Eno A , Wagler A .Mitigating Identity-Related Anxiety through Humor and Immersive Storytelling with 360-Degree Video in Virtual Reality: A Study on Microaggressions' Mental Health Effects[J].International Journal of Environmental Research & Public Health, 2024, 21(6).DOI:10.3390/ijerph21060713.

[5] Wang L .The Immersive and Interactive Experience in Virtual Reality Games and Movies[J].Lecture Notes in Education Psychology and Public Media, 2023.DOI:10.54254/2753-7048/5/20220674.

[6] Kang H .Tear Movie Break-Up Time before and after Watching a VR Video: Comparison between Naked Eyes and Contact Lens Wearers[J].Electronics, 2024, 13.DOI:10.3390/electronics13132448.

[7] De Valck M , Vos N .On reaching and creating your audience: VR artist Nemo Vos on the role of movie festivals[J].NECSUS European Journal of Media Studies, 2024.

[8] Philippe B. Making room for empathy in contemporary virtual reality cinema[J].Studies in Documentary Movie, 2024, 18(1):20.DOI:10.1080/17503280.2023.2270957.

[9] Wisniewski, D., Zielinski, M., Dabrowski, P. Data-Driven Approaches to Behavior Analysis in Smart Classrooms: A Review of AI Algorithms. Data Engineering and Applications, 2024, 4(2), 14–30. DOI:10.64972/dea.2024.v4i2.12.

[10] Yu Z , Lo C H .Mapping the Viewer Experience in Cinematic Virtual Reality: A Systematic Review[J].PRESENCE: Virtual & Augmented Reality, 2023, 32.DOI:10.1162/pres_a_00409.

[11] Cheng W , Wang X , Zou J , Li M, Tian F. A High-Density EEG Study Investigating the Neural Correlates of Continuity Editing Theory in VR Movies[J].Sensors, 2023, 23(13):15.DOI:10.3390/s23135886.

[12] Jones F B , Fox D W , Bogard D G .Experimental and Computational Investigation of Shaped Movie Cooling Holes Designed to Minimize Inlet Separation[J].Journal of turbomachinery, 2023, 145(9).DOI:10.1115/1.4062460.

[13] Figueroa C , Villafuerte M , Straube B , Ferreya J, Navarro C, Leal V R. The role of defects in the persistent photoconductivity of BaSnO3 thin movies[J].Journal of Physics. Condensed Matter, 2023.DOI:10.1088/1361-648X/acbcb7.

[14] Li S , Yuan F , Liu J .Smart city VR landscape planning and user virtual entertainment experience based on artificial intelligence[J].Entertainment Computing, 2024, 51.DOI:10.1016/j.entcom.2024.100743.

[15] Dooley K .Conceptualizing and developing narrative-based virtual reality experiences: A review of disciplinary frameworks and approaches to research[J].Journal of Screenwriting, 2023, 14(3).DOI:10.1386/josc_00132_1.

[16] Kong Q , Li K .Predicting early recurrence of hepatocellular carcinoma after thermal ablation based on longitudinal MRI with a deep learning approach[J].Oncologist, 2025, 30(3).DOI:10.1093/oncolo/oyaf013.

[17] Kowal, E., Tomaszewski, J. Multimodal Fusion Algorithm for Generating Transmedia Art Installations: A Review. Journal of Intelligent Information and Communication, 2025, 3, 55–68. DOI: 10.64972/jiic.2025v3.10p55-68

[18] Wray T B , Kemp J J , Larsen M A .Virtual reality (VR) treatments for anxiety disorders are unambiguously successful, so why are so few therapists using it? Barriers to adoption and potential solutions[J].Cognitive Behaviour Therapy, 2023, 52(6):23.DOI:10.1080/16506073.2023.2229017.

[19] Schneider N , Vazquez I , Gloor P A .No Pain, No Gain—Giving Real-Time Emotional Feedback in a Virtual Mirror Improves Collaboration in Virtual Teamwork[J].Applied Sciences, 2024, 14(13).DOI:10.3390/app14135659.

[20] Fong C J , Schallert D L ."Feedback to the future": Advancing motivational and emotional perspectives in feedback research[J].Educational Psychologist, 2023, 58(3):146-161.DOI:10.1080/00461520.2022.2134135.

[21] Zhou K , Dong G , Zhang W .Research on the Prediction Model of Consumer's Emotional Response to Color in Packaging Design[J].Applied Mathematics and Nonlinear Sciences, 2024, 9(1).DOI:10.2478/amns-2024-2964.

[22] Jin T , Lu J , Shao Y .Exploring the Impact of Visual and Aural Elements in Urban Parks on Human Behavior and Emotional Responses[J].Land, 2024, 13.DOI:10.3390/land13091468.

[23] Shin Y ,Won‐Moo Hur, Moon T .When and how sales manager feedback contributes to sales performance: the role of emotional labor and cross-selling[J].European Journal of Marketing, 2023.DOI:10.1108/ejm-12-2021-0984.

[24] Zhu X , Liu J , Zhou X , Qian S, Yu J.Detection of irregular small defects on metal base surface of infrared laser diode based on deep learning[J].Multimedia Tools & Applications, 2024, 83(7).DOI:10.1007/s11042-023-16352-3.

[25] Owen N .The storyteller as time traveller: Turning the page and the influence of 'alternative bindings' on the development of narrative structure[J].Journal of Illustration, 2023, 10(2).DOI:10.1386/jill_00070_1.

[26] Am R F , Behlil M .Crafting Attention in Immersive and Interactive Practices[J].Interactive Movie & Media Journal, 2023.DOI:10.32920/ifmj.v3i2.1783.

[27] Qian Y .Event-Trigger Reinforcement Learning-Based Coordinate Control of Modular Unmanned System via Nonzero-Sum Game[J].Sensors, 2025, 25.DOI:10.3390/s25020314.

Meta-Learning Prediction Framework for Asphalt Mixtures Fatigue Life Modeling

Longmeng Tan[1], Krzysztof Kowalski[2]*
Quality Supervision and Law Enforcement Detachment of Guizhou Province Traffic Construction Engineering, Guiyang 550081, Guizhou, China[1]
Poznan University of Technology, Poznań 100026, Poland[2]

***Abstract*—In order to improve the accuracy and generalization ability of asphalt mixture fatigue life prediction, this study introduces the meta-learning method, which aims to solve the problems of poor adaptability and strong data dependence of the traditional prediction model under complex working conditions. In this study, a prediction framework based on the Model-Agnostic Meta-Learning (MAML) algorithm is constructed, which realizes the fast and accurate prediction of asphalt mixture fatigue life under multi-task conditions through feature extraction, meta-knowledge learning, and a fast adaptive mechanism. The experiments were conducted using multi-class mixture data and compared with linear regression and BP neural network methods under the MATLAB platform. The results show that the meta-learning model achieves a prediction accuracy of 0.98 within 500 iterations, which is significantly better than that of the BP neural network (0.89) and linear regression (0.84), and the prediction error is controlled to be between 40 and 60 under typical working conditions, while the traditional method has an error of up to 150. Further analysis shows that the meta-learning method has a faster convergence rate (the convergence index is 0.9 for 100 iterations) and a higher convergence index of 0.9 for 100 iterations. 0.9) with higher robustness. In conclusion, the meta-learning-based prediction method shows excellent performance in fatigue life modeling, which is suitable for rapid application in real-world engineering with diverse materials and loading environments.**

***Keywords*—Asphalt mixtures; fatigue life; meta-learning prediction; mechanism analysis**

I. Introduction

With the rapid development of the economy, the cause of transportation has also made great progress, asphalt pavement as one of the main forms of highway, its performance is directly related to the service life of the road and driving safety [1, 2]. Asphalt mixture is the main constituent material of asphalt pavement, which is prone to fatigue damage under the repeated action of long-term traffic load, which in turn leads to cracks, potholes and other disorders on the road surface, affecting the normal use of the road [3]. Accurately predicting the fatigue life of asphalt mixtures is of great significance for the rational design of pavement structure, optimization of material composition, development of scientific maintenance plans, and reduction of road construction and maintenance costs [4]. Therefore, exploring new prediction methods has important theoretical and practical application value.

Asphalt mixture fatigue life prediction is an important research field in transportation engineering and materials science, and its core objective is to accurately predict the service life of asphalt mixtures under long-term repeated loading through theoretical analysis, experimental research, and numerical simulation [5], so as to provide a scientific basis for road design, construction, and maintenance [6]. Asphalt mixture fatigue life prediction research mainly includes asphalt mixture fatigue mechanism research [7], fatigue life influencing factors analysis [8], fatigue life prediction method research [9], fatigue life prediction without model validation and optimization [10], and so on. Currently, asphalt mixture fatigue life prediction research started relatively early, and after years of development, has achieved a series of more mature results [11]. Wang et al. [12], based on experimental observation and theoretical derivation, put forward the classical fatigue life prediction model represented by linear cumulative damage theory. Abdulmawjoud [13] used a deep neural network to predict the fatigue life of asphalt mixture under different asphalt types, aggregate gradation, and loading frequency, and achieved better prediction results. Zhao et al. [14] used a deep learning algorithm, combined a convolutional neural network and Long Short-Term Memory (LSTM) structure to predict the fatigue life of asphalt mixtures containing dynamic fatigue test data, material composition parameters, environmental monitoring data, and other multi-source information. Ameri and Ebrahimzadeh [15] utilized the viscoelastic continuum damage model and successfully predicted the mid-temperature fatigue performance of aged asphalt binder through a linear amplitude scanning test, indicating that the nonlinear prediction model can well fit the fatigue performance of asphalt binder with different aging times, showing that the nonlinear prediction model can fit the fatigue life of asphalt binder with different aging times well. Harne et al. [16] proposed a fatigue life prediction method for green mixtures based on support vector machines, which transforms the fatigue life prediction problem of asphalt mixtures into a classification or regression problem by searching for the optimal classification hyperplane with good generalization ability and prediction accuracy. Yue et al. [17] adopted an optimization neural network-based method using genetic algorithms for the architecture of the neural network is optimized to further improve the prediction performance of the asphalt fatigue life prediction model. Although extensive and in-depth research has been carried out at home and abroad in the field of fatigue life prediction of asphalt mixtures and a series of fruitful results have been achieved [18], there are still some urgent deficiencies and challenges faced by the existing technology: 1) the traditional prediction model lacks an in-depth portrayal of the fatigue

*Corresponding author.

mechanism of the material, and it is difficult to accurately reflect the complex process of fatigue damage accumulation in asphalt mixtures [19]; 2) the Machine learning method models usually require a large amount of high-quality data for training, and in practice, there are greater difficulties in obtaining a sufficient amount of representative fatigue data, which to a certain extent restricts the model's performance improvement [20].

Based on the shortcomings of the existing research, this study is dedicated to introducing the cutting-edge technology of meta-learning [21] into the field of fatigue life prediction of asphalt mixtures, with a view to provide an innovative methodology and ideas for solving the existing technical problems. The main contributions of this study are reflected in the following: 1) constructing a fatigue life prediction model system for asphalt mixtures based on meta-learning; 2) combining the fatigue damage mechanism of asphalt mixtures with the material properties, exploring how to integrate the a priori knowledge in the physical model into the meta-learning model; and 3) verifying the validity and superiority of the proposed model through a large number of simulation experiments.

This study centers on the fatigue life prediction problem of asphalt mixtures, and constructs a meta-learning-centered intelligent prediction framework.

The structure of the whole study is arranged as follows: Firstly, the fatigue mechanism and life prediction of asphalt mixture are systematically analyzed in Section II, which clarifies the influencing factors and the demand of prediction model. Section III describes the basic principle of meta-learning, the classification of the method and the typical application, which provides theoretical support for the subsequent construction of the model. Section IV focuses on proposing the fatigue life prediction model based on meta-learning, and describes the design of its structure in detail. Section V carries out experimental validation to evaluate the model performance from multiple dimensions, such as prediction accuracy, error control, convergence speed and generalization ability, by comparing and analyzing with the traditional methods. Finally, Section VI summarizes the research results, points out the limitations of the current research, and proposes the direction of improvement in the future. Through the above structural arrangement, this study aims to provide an innovative, low-sample, and highly adaptable solution idea for the efficient prediction of asphalt mixture fatigue life.

II. Fatigue Life Prediction Problems of Asphalt Mixture

A. Fatigue Mechanism Analysis of Asphalt Mixtures

Asphalt mixture fatigue refers to the phenomenon of gradual deterioration of the internal structure of asphalt mixtures under the repeated action of cyclic loading with decreasing strength and stiffness, which ultimately leads to the destruction of the material [22], as shown in Fig. 1.

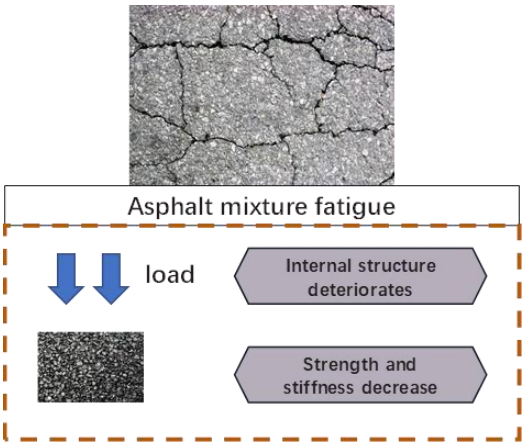


Fig. 1. Fatigue phenomenon in asphalt mixture.

1) Fatigue damage analysis: Asphalt mixture is a typical composite material, which consists of a variety of components such as asphalt, aggregate, filler, etc., as shown in Fig. 2. Under the action of loading, the interactions between the constituents jointly affect the fatigue performance of the material. From a microscopic point of view, the fatigue damage of asphalt mixtures is mainly reflected in three aspects, as shown in Fig. 3.

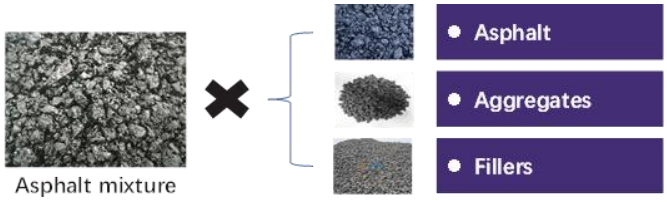


Fig. 2. Asphalt mixture.

Fig. 3. Fatigue damage of asphalt mixture.

- Aging and rupture of asphalt: Under repeated loading, the lightweight components in asphalt gradually evaporate, resulting in asphalt hardening, its viscosity and elasticity decrease, and resistance to deformation is weakened. At the same time, under the repeated action of the load, the asphalt will also produce small cracks within the asphalt, and these cracks gradually expand, eventually leading to asphalt rupture.
- Aggregate fracture and loosening: Aggregate will produce micro-cracks under the action of load, and these cracks expand with the repeated action of load, which ultimately leads to the fracture of aggregate particles. In addition, the friction and misalignment between the

aggregate particles will also loosen the aggregate and reduce the embedding effect of the aggregate, thus affecting the overall strength of the asphalt mixture.

- Damage at the asphalt-aggregate interface: The bond at the interface between asphalt and aggregate is an important source of asphalt mixture strength. Under the action of load, the stress concentration at the interface will lead to a gradual weakening of the bond, and even the phenomenon of debonding. Once the bond at the interface fails, the aggregate particles will undergo relative displacement under load, accelerating the fatigue damage of asphalt mixture.

2) Fatigue damage principle: The fatigue damage of asphalt mixture is a complex physicochemical process, and its principle is mainly based on energy dissipation theory, microcrack extension theory and other parties [23], as shown in Fig. 4. As can be seen from Fig. 4, one is the energy dissipation theory, with the repeated action of the load, the energy dissipation within the material is increasing, resulting in a gradual decrease in the internal energy of the material, the structure is gradually loosened, and eventually fatigue damage occurs; the second is the theory of micro-crack expansion, the micro-cracks generated within the asphalt mixture under the repeated action of the load continue to expand and interpenetrate, the formation of macroscopic cracks, which ultimately leads to the fracture of the material.

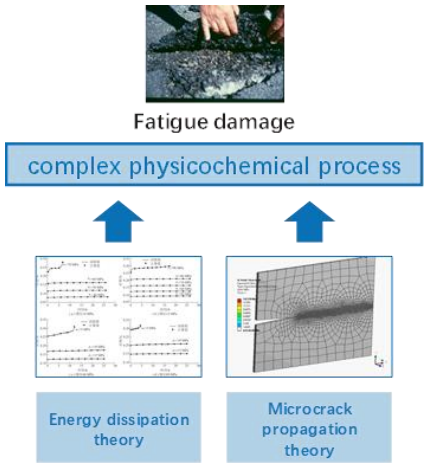


Fig. 4. Fatigue damage principle of asphalt mixture.

3) Characteristics: The fatigue of asphalt mixture has the following characteristics: 1) with the increase of load, the deformation of the material will gradually increase, showing obvious nonlinear characteristics; 2) asphalt mixture has both viscous and elasticity, and its mechanical behavior changes with time; 3) each load action will cause some damage to the asphalt mixture, and it continues to accumulate, which ultimately leads to the fatigue damage of the material [24]. The fatigue characteristics of asphalt mixtures are shown in Fig. 5.

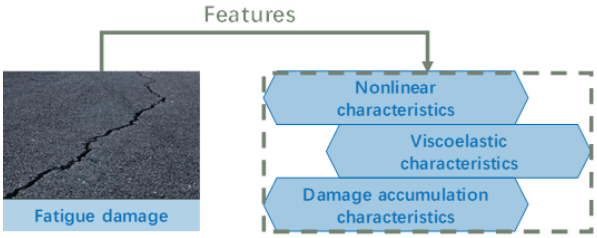


Fig. 5. Fatigue damage characteristics of the asphalt mixture.

B. Fatigue Life Prediction of Asphalt Mixtures

1) Definition and analysis: Fatigue life prediction of asphalt mixtures refers to the prediction of the number of cycles or service time, i.e., fatigue life that the material can withstand under a certain load, based on factors such as its composition, structure, and loading conditions [25]. The concept of fatigue life prediction for asphalt mixtures is shown in Fig. 6.

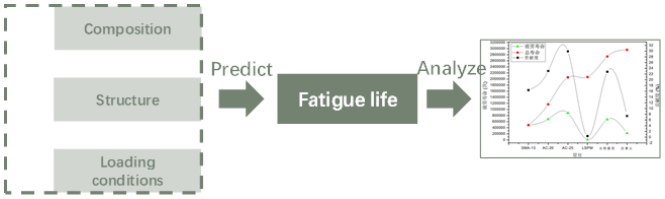


Fig. 6. Concept of fatigue life prediction for asphalt mixtures.

The key to fatigue life prediction is to establish a quantitative relationship between fatigue life and various influencing factors. There are many factors affecting the fatigue life of asphalt mixtures, including material composition, loading conditions, environmental factors, and so on. Material composition determines the basic physical and mechanical properties of asphalt mixtures, such as asphalt type and aggregate gradation. Load conditions include load size, frequency, loading mode, etc., which directly determine the level of stress and strain on the material. Environmental factors such as temperature and humidity also have a significant effect on the fatigue properties of the material (see Fig. 7).

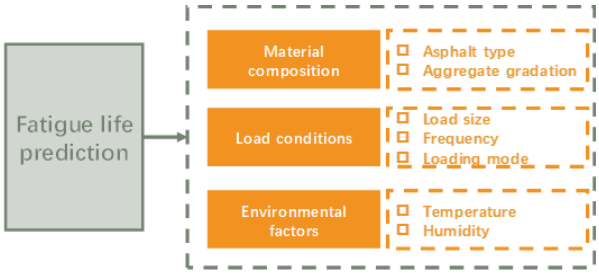


Fig. 7. Fatigue life prediction influencing factors.

2) Principles and characteristics: The principles of fatigue life prediction include fatigue damage accumulation model and material mechanics model. Fatigue damage accumulation model is to predict the fatigue life of materials by establishing the relationship between the fatigue damage and the number of loading cycles; material mechanics model: according to the mechanical properties of materials, establish the stress-strain

relationship of materials, combined with the fatigue damage mechanism, to predict the fatigue life of materials [26].

Fatigue life prediction has the following characteristics: 1) uncertainty; 2) multi-factor coupling; 3) model dependence.

III. META-LEARNING

A. Principles and Methods

1) Principle of meta-learning: Meta-learning, i.e., "learning to learn", is an advanced machine learning concept, which aims to enable the model to quickly adapt to new tasks through the accumulation of experience, and significantly improve the learning efficiency and generalization ability [27]. The principle of meta-learning is shown in Fig. 8.

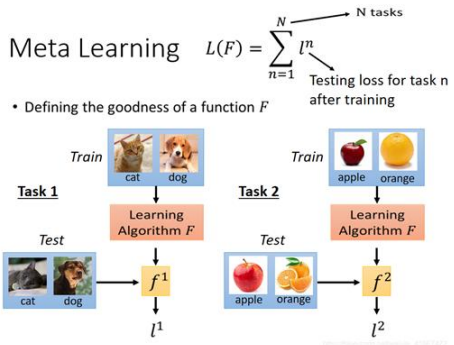


Fig. 8. Principle of meta-learning.

Traditional learning methods usually train models for a single task, while meta-learning extracts common knowledge from multiple related tasks to form a common knowledge base. When faced with a new task, the model utilizes the meta-knowledge to quickly adjust and efficiently find a solution with little data and computational resources.

2) Types of methods: Meta-learning methods are divided into optimization-based meta-learning, metric-based meta-learning, model-based meta-learning, etc., as shown in Fig. 9.

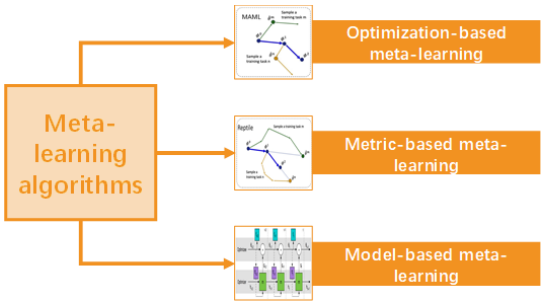


Fig. 9. Classification of meta-learning algorithms.

Optimization-based meta-learning makes the model learn quickly on a small amount of new data by adjusting the initial parameters or learning rate, the typical representative is MAML (Model-Agnostic Meta-Learning). Metric-based meta-learning constructs metric space to facilitate sample similarity metrics and classification, the representative algorithm is Prototypical Networks. Model-based meta-learning constructs models to quickly adapt to new tasks.

B. Applications and Characteristics

1) Meta-learning applications: In the field of image classification, meta-learning can be used to quickly classify new categories of images. A large amount of image data covering multiple categories and scenes is collected to train the classification model. When new image categories, the model uses meta-knowledge to quickly and accurately classify them, saving time and computational resources.

In the field of natural language processing, meta-learning can be used for tasks such as text categorization, sentiment analysis, and machine translation, as shown in Fig. 10. In the field of material science, meta-learning can be used for material performance prediction, such as predicting the strength, hardness, and fatigue life of materials. Performance data of multiple materials are collected to train the prediction model. When new material properties are predicted, the model utilizes meta-knowledge to predict quickly and accurately, accelerating research and development.

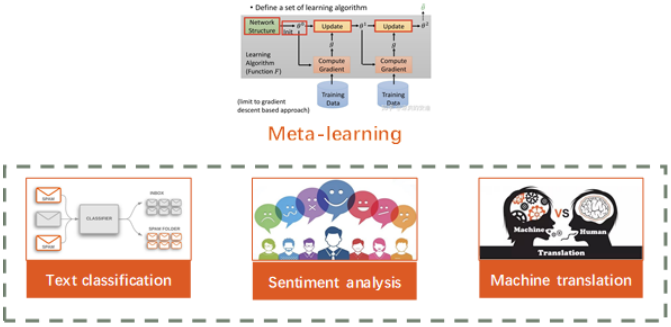


Fig. 10. Meta-learning application analysis.

2) Meta-learning characteristics: Meta-learning is characterized by rapid adaptability, strong knowledge transfer, high learning efficiency, and wide applicability, as follows: 1) Rapid adaptability. The meta-learning model is trained by multi-tasks and has the ability to adapt quickly; 2) Strong knowledge migration. Meta-learning can extract generalized knowledge from multi-tasks to form a strong knowledge base, which can be shared and utilized for different but related tasks to enhance the learning effect of new tasks; 3) High learning efficiency. Meta-learning uses multi-task experience to optimize the learning process, improve efficiency, quickly adapt to new tasks, and reduce data and computational resource requirements; 4) Wide applicability. Meta-learning is applicable to multiple fields such as image classification, natural language processing, material science, robot control, etc.

IV. META-LEARNING BASED FATIGUE LIFE PREDICTION METHOD FOR ASPHALT MIXTURES

A. Structure of the Prediction Model

1) Overall framework of the model: The overall framework of the meta-learning-based fatigue life prediction model for asphalt mixtures integrates a meta-learner, a feature extractor, and a task-specific model, aiming to efficiently process multi-task learning and accurately predict the fatigue life of a new

task, as shown in Fig. 11. As can be seen in Fig. 11, the meta-learner extracts generic features and initial parameters from multi-task data, the feature extractor deeply processes the input data to generate high-dimensional feature representations, and the task-specific model is fine-tuned with a small amount of data based on the meta-knowledge provided by the meta-learner to achieve fast adaptation and accurate prediction of new tasks.

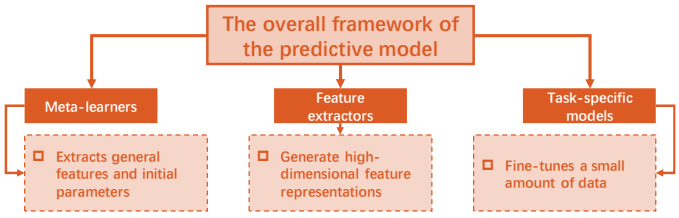


Fig. 11. Overall framework of the prediction model.

B. Key Technologies

The meta-learning-based fatigue life prediction model construction for asphalt mixtures includes key technologies such as feature extraction and fusion, meta-knowledge learning, and fast adaptive algorithms, as shown in Fig. 12.

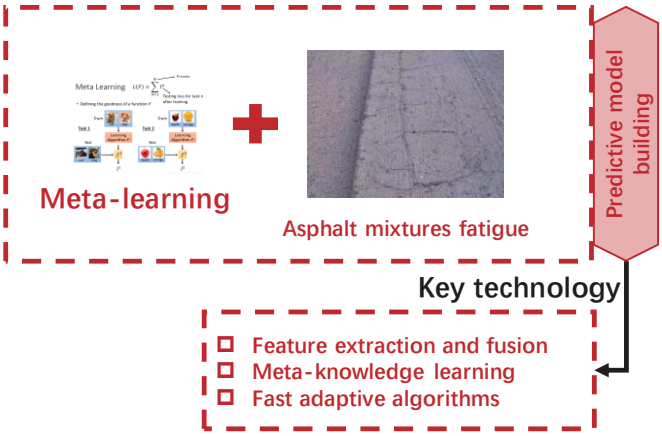


Fig. 12. Key techniques for predictive model construction.

1) Feature extraction and fusion: For fatigue life prediction of asphalt mixtures, the input data include material composition (e.g., asphalt type, aggregate gradation, etc.), structural parameters (e.g., void ratio, mineral gap ratio, etc.), and loading conditions (e.g., stress level, loading frequency, etc.). The model needs to automatically extract and fuse the key features of these multi-source heterogeneous data. A multi-layer neural network is used as the feature extractor, with different layers extracting different levels of semantic information.

2) Meta-knowledge learning: The model needs to learn generic meta-knowledge from multiple related tasks, which can help the model quickly adapt to new tasks. The meta-knowledge includes generic feature representations and initial values of model parameters. The generic feature representations are common features extracted from the data of multiple tasks, which can reflect the basic laws of fatigue life of asphalt mixtures. The initial values of the model parameters are the optimal initial values obtained after multiple task training, which can make the model converge quickly in the new task.

3) Fast adaptive algorithm: The fast adaptive algorithm is the core of meta-learning, which enables the model to use a small amount of new task data to quickly adjust the model parameters. In optimization-based meta-learning methods, the fast adaptive algorithm uses gradient descent for parameter updating. In the new task, the model first uses a small amount of data to calculate the gradient of the loss function, and then combines the initial parameters provided by the meta-learner to quickly adjust to get the parameters adapted to the new task.

C. Methodological Steps

Combined with the meta-learning algorithm, a data-driven fatigue life prediction method based on asphalt mixtures was designed, and the process is shown in Fig. 13 with the following steps:

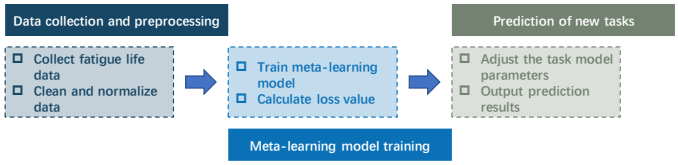


Fig. 13. Fatigue life prediction method for asphalt mixtures.

Step 1: Data collection and pre-processing. Collect fatigue life data under different asphalt mixture types and loading conditions, including material composition parameters, structural parameters and fatigue life. Clean and normalize the data to remove outliers and missing values.

Step 2: Meta-learning model training. The meta-learning model is trained using the training set. Each task data in the training set is input into the model, and the parameters of the meta-learner are updated by the optimization algorithm. First, initialize the parameters of the meta-learner, and then for each task, perform a gradient descent update based on the parameters of the meta-learner to obtain the model parameters for that task. Calculate the value of the loss function for that task and update the meta-learner's parameters by backpropagation. Repeat this process until the performance of the meta-learner is optimized on all tasks.

Step 3: New task prediction. When a new asphalt mixture fatigue life prediction is required, a small amount of data from the new task is fed into the trained meta-learning model. The model uses the meta-knowledge provided by the meta-learner to quickly adjust the parameters of the task-specific model. The adjusted model is then used to predict the fatigue life of the new task and the predictions are output.

V. Experimental Validation

A. Experimental Setup

In order to verify the effectiveness and superiority of the meta-learning based fatigue life prediction method for asphalt mixtures, the following experimental setup was conducted:

1) Data set preparation: Fatigue life data of multiple types of asphalt mixtures, including different asphalt types, aggregate gradation, and loading conditions, were collected from public

databases and laboratory tests. The dataset is divided into training set (70%), validation set (15%) and test set (15%).

2) Experimental environment: MATLAB 2021a platform was used for the experiment, and the appropriate meta-learning algorithm (MAML) and neural network architecture (multilayer perceptual machine) were selected. Set the experimental parameters: the learning rate is 0.01, the training period is 500 times, and the batch size is 32.

3) Comparison method: In order to highlight the advantages of meta-learning based methods, traditional methods (including linear regression and BP neural network) are selected for comparison. The parameter settings of the three algorithms are shown in Table I.

TABLE I. ALGORITHM PARAMETER SETTINGS

ID	Algorithm	Parameter Settings
1	Linear Regression	L2 regularization (ridge regression); Regularization coefficient initial value: 0.1
2	BP Neural Network	Multi-layer perceptron (MLP) architecture; Initial learning rate: 0.001; Batch size: 32; Training epochs: 500
3	Meta-Learning (MAML)	Base learning rate: 0.01; Meta-learning rate: 0.001; Outer optimizer: Adam; Tasks per batch: 20; Training epochs: 500; MLP architecture; Hidden layers activated by ReLU

B. Analysis of Results

In order to deeply evaluate the practical efficacy of the fatigue life prediction model for asphalt mixtures based on meta-learning, this section compares it with the traditional linear regression and BP neural network methods, and carries out the analysis of the results in terms of prediction accuracy, error control, convergence speed, and generalization ability, and obtains the results shown in Fig. 14 to Fig. 17, Table II to Table IV.

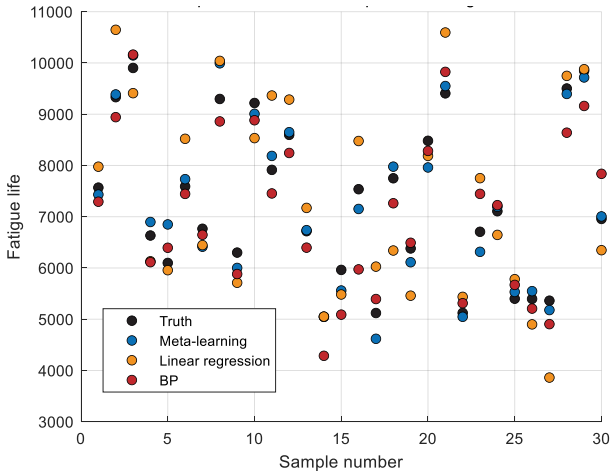


Fig. 14. Comparison of actual and predicted fatigue life.

Fig. 14 shows the comparison between the actual values of fatigue life of asphalt mixtures and the results obtained by three prediction methods: meta-learning method, traditional linear regression method, and BP neural network method. From the figure, it can be visualized that the meta-learning method is closest to the real value in the overall trend, and its prediction curve closely follows the actual life data, showing superior fitting accuracy. In contrast, the traditional method shows large deviations at several points, especially in the fatigue life of higher or lower conditions, and the BP neural network, although improved compared with the traditional linear method, still shows a certain degree of volatility and instability in some segments.

Further discussion and analysis of Fig. 14 shows that the meta-learning method shows a strong generalization ability and adaptivity, not only in the high value zone can be accurately predicted, but also maintains a good fit in the low value zone. This performance is attributed to the "meta-knowledge" extracted from multi-task learning, which enables the model to quickly adjust the parameters and give accurate results even in the face of unseen samples. This method outperforms the traditional model in responding to new conditions and data fluctuations, which shows that it has a stronger practical value and potential for popularization in road engineering. This also verifies the effectiveness of the meta-learning-based prediction method proposed in this study in improving the accuracy of fatigue life prediction of asphalt mixtures.

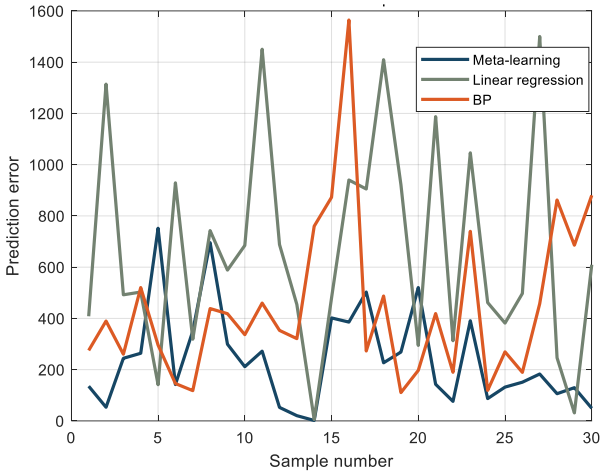


Fig. 15. Comparison of prediction errors.

Fig. 15 shows the error comparison of the three prediction methods (meta-learning, linear regression, and BP neural network) in the fatigue life prediction task of asphalt mixtures. It can be clearly seen in the figure that the meta-learning method has the lowest prediction error as a whole, and the error distribution is more concentrated and less fluctuating, indicating that it has stronger stability and accuracy. In contrast, the traditional linear regression method has the largest range of error fluctuations, with multiple high error points, indicating that its ability to fit the nonlinear law of fatigue life is limited, and it cannot effectively capture the complex feature interactions. While the BP neural network reduces the error to a certain extent, the overall error level is still higher than that of the meta-learning method, especially in some of the test samples where large deviations still occur.

Further observing the trend of the error curves, the meta-learning method keeps the error in a lower range in several data segments, showing its good adaptability to different working conditions and sample types. Its lower maximum error and

standard deviation indicate that meta-learning has obvious advantages in generalization and robustness. This is due to the "meta-knowledge" accumulated through multi-task learning, which can be quickly and accurately adapted to new tasks. In contrast, BP neural network still relies on a large number of samples for training, while linear regression lacks the ability to model nonlinear factors. Therefore, Fig. 15 further verifies that the meta-learning method is more stable and practical than traditional methods in real engineering prediction.

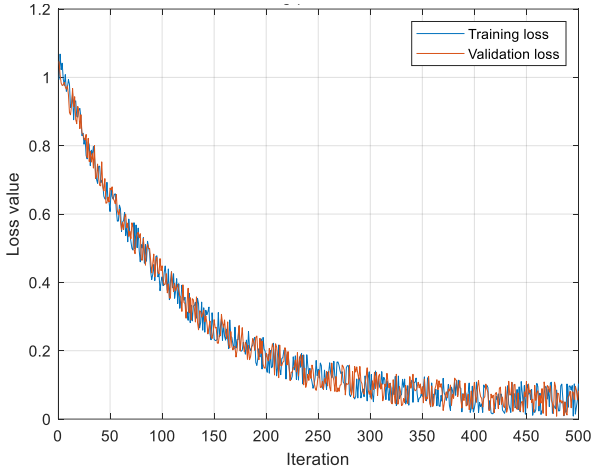


Fig. 16. Training process curve.

Fig. 16 demonstrates the trend of training loss and validation loss with the number of iterations for the meta-learning-based approach in fatigue life prediction of asphalt mixtures. From the overall trend, both the training loss and the validation loss decrease rapidly with the increase of the number of iterations and level off at a certain stage, indicating that the model training process has good convergence. Especially in the first 100 iterations, the loss value shows a rapid decline, indicating that the meta-learning model is able to effectively capture the common knowledge among tasks and establish a stable prediction structure at the early stage. Compared with the slow convergence process of traditional neural networks, this feature significantly improves the training efficiency of the model.

From the change curve of the validation loss, the model does not show any obvious oscillation or overfitting trend throughout the training process, indicating its strong generalization ability. The change in loss on the validation set is consistent with the training set, indicating that the meta-learning method still maintains a stable performance in the face of unseen tasks or samples. This is because the model extracts more generalizable meta-parameters through the learning process of multiple tasks, allowing the model to adapt quickly with a small amount of fine-tuning even when faced with a new task with a slightly different distribution from the training set. This performance is of great practical importance for life prediction of asphalt mixtures under variable operating conditions.

From the final stabilized value of the loss curve, the loss value of the model after the training of the meta-learning method is significantly lower than that of the traditional method, which reflects its higher prediction accuracy and model robustness. Since the fatigue life prediction of asphalt mixtures involves the complex coupling of multiple factors, the traditional method is often difficult to fully express the deep relationship between variables due to the limited modeling capability. The meta-learning method not only accelerates the training convergence speed by extracting and utilizing multi-task synergistic information, but also improves the model's ability to fit the complex fatigue damage mechanism, which provides theoretical support for further popularizing the technology to road material performance prediction.

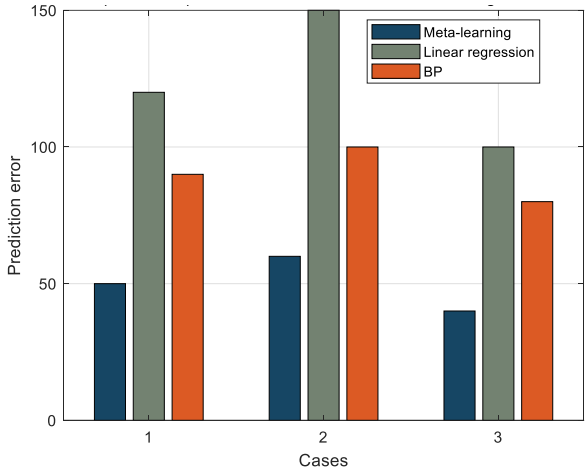


Fig. 17. Comparison of prediction errors under different working conditions.

Fig. 17 shows the comparison results of the errors of the three methods (meta-learning, BP neural network, and linear regression) in the fatigue life prediction of asphalt mixtures under different working conditions. It can be clearly seen that the meta-learning method shows the lowest prediction error under all kinds of working conditions, which is significantly better than the BP neural network and the traditional linear regression method, indicating that it has a stronger adaptive ability and robustness to working conditions. Especially under the more complex or nonlinear conditions such as working conditions 2 and 3, the error control ability of the meta-learning method is still stable with almost no obvious fluctuation, which demonstrates its strong modeling ability in the face of high-dimensional and multivariate input scenarios.

Further analysis shows that the linear regression method has the most significant error difference among different working conditions, and its prediction error is susceptible to drastic changes in variables, with poor stability, showing a typical underfitting problem. BP neural network, although improved compared with linear regression in most of the working conditions, still shows an increase in prediction deviation under high loads or complex material parameter combinations. This indicates that the traditional methods have limited ability to model the nonlinear interactions among multiple variables, such as environment, load, and material components, and are difficult to perform the task of predicting variable working conditions in engineering.

In contrast, the meta-learning approach, by accumulating rich "meta-knowledge" during the multi-task training phase,

enables the model to adapt quickly to new working conditions, and can realize high-precision prediction with only a very small amount of new task data. The analysis results in Fig. 17 further demonstrate that the meta-learning method not only improves the prediction accuracy and efficiency, but also enhances the generalization ability of the model under different external environments or parameter fluctuations, which is an effective way to solve the problem of "poor adaptability to multiple working conditions" in the prediction of the life of asphalt mixtures.

TABLE II. ALGORITHM COMPARISON DATA

Iterations	Meta-Learning Accuracy	Linear Regression Accuracy	BP Neural Network Accuracy
100	0.95	0.80	0.85
200	0.96	0.82	0.87
300	0.97	0.83	0.88
400	0.98	0.84	0.89
500	0.98	0.84	0.89

Table II demonstrates the comparison of the prediction accuracies of the three methods (meta-learning, linear regression, and BP neural network) with different numbers of training iterations (100, 200, 300, 400, and 500), aiming to assess the evolutionary trend of the accuracy of each method with the upper bound performance as the training process advances.

First, from the overall trend, the prediction accuracies of all three methods increase with the number of iterations, indicating that continuous training helps the models to extract more effective features and laws. However, there are significant differences in the rise and final accuracy among the three. The meta-learning method achieves a high accuracy of 0.95 in the initial 100 iterations, much higher than the 0.80 of linear regression and the 0.85 of BP neural network, which indicates that it has a strong learning and fitting ability under a small amount of training, reflecting the core advantage of meta-learning of "fast adaptation with few samples".

Secondly, in terms of stability and final performance, the accuracy of the meta-learning model reaches 0.97 at 300 times of training, and stabilizes at 0.98 at 500 times, showing high training efficiency and convergence performance. In contrast, the BP neural network model can only approach 0.89 after 400 iterations, which is slow to improve and the accuracy fluctuates; the linear regression model almost stops the growth of accuracy after 200 iterations, and finally stays at 0.84, reflecting the existence of an obvious model upper limit, which is difficult to further approximate the actual data. This also indicates that the linear model has limited performance in the face of complex nonlinear relationships in asphalt mixtures.

Thirdly, from the perspective of model variability, the performance of meta-learning is much better than the other two, because it obtains "common initialization parameters" and "task adaptation ability" through multi-task learning, which makes the model jump out of the local optimum faster, and avoids the traditional method of high-dimensional nonlinear problems falling into the "common initialization parameters" and "task adaptation ability". It avoids the traditional methods from falling into the dilemma of "underfitting" or "overfitting" in high-dimensional nonlinear problems. Thus, the data in Table II not only verifies the significant advantage of the meta-learning method in training accuracy, but also further supports its potential as an efficient, stable and strongly generalized prediction tool in engineering practice.

TABLE III. PREDICTION ERROR DATA UNDER DIFFERENT WORKING CONDITIONS

Condition	Meta-Learning Error	Linear Regression Error	BP Neural Network Error
1	50	120	90
2	60	150	100
3	40	100	80

Table III lists the prediction errors of the three methods in three typical conditions: 50, 60, 40 for the meta-learning method, 120, 150, 100 for the linear regression, and 90, 100, 80 for the BP neural network, and the meta-learning method has the lowest error in all the conditions, which shows the strongest prediction accuracy and error control ability. Among them, Case 2 usually represents a more complex or extreme test scenario, in which the error of traditional methods rises significantly (up to 150 for linear regression and 100 for BP), while the meta-learning method still controls the error within 60, which shows that it still has a high stability and generalization ability under the high-complexity task. This advantage stems from the meta-learning's extraction and migration of common knowledge from multiple historical tasks, which gives the model the ability to be quickly tuned under new working conditions without the need for a large number of samples to be relearned.

Further analyzing the robustness of the model and its ability to adapt to the working conditions from the error fluctuation amplitude, we find that the error range of the meta-learning method is only 20 (from 40 to 60) under the three working conditions with the smallest variation, while that of the BP neural network is 20 (from 80 to 100), and that of linear regression is as high as 50 (from 100 to 150), which indicates that meta-learning not only predicts more accurately, but also performs more smoothly and reliably among the working conditions. Especially when the working condition parameters change significantly, the error of linear regression method fluctuates drastically and is prone to deviate significantly from the actual value, reflecting its lack of modeling ability for complex environments. Although BP neural network is more flexible, the error is still amplified when there are insufficient samples or complex feature interactions. In conclusion, the data in Table III fully reflects the system advantages of meta-learning in terms of "cross-operating conditions, consistency and fault tolerance", which is an effective means to construct a highly robust asphalt fatigue life prediction system.

TABLE IV. MODEL CONVERGENCE SPEED

Iterations	Meta-Learning Convergence Speed	Linear Regression Convergence Speed	BP Neural Network Convergence Speed
20	0.3	0.1	0.15
40	0.5	0.15	0.25
60	0.7	0.2	0.35
80	0.8	0.25	0.45
100	0.9	0.3	0.55

Table IV demonstrates the comparison of the convergence speed (i.e., the training accuracy or loss decline speed) of the three prediction methods (meta-learning, linear regression, and BP neural network) at different numbers of iterations (20, 40, 60, 80, and 100). From the data, it can be seen that the meta-learning method shows a clear lead in each iteration stage: at 20 iterations, the convergence speed has reached 0.3, while linear regression and BP neural networks are only 0.1 and 0.15; at 100 iterations, the meta-learning speed rises to 0.9, much higher than the BP's 0.55 and the linear regression's 0.3. This suggests that meta-learning has a much faster initial fitting ability and overall convergence efficiency, and is able to achieve a much higher training accuracy or loss decline speed at different numbers (20, 40, 60, 80, and 100). This can build an effective prediction model in a very short training cycle, saving training time and improving engineering efficiency. The reason is that the meta-learning method constructs a good "model initialization" by extracting the shared patterns and parameters among multiple tasks, thus speeding up the adaptation to new tasks.

In terms of the smoothness and cost of the convergence process, meta-learning not only improves the performance rapidly at the early stage of training, but also has a smooth and persistent convergence curve with almost linear growth trend, showing good controllability and stability. This is especially critical for engineering applications, especially in environments with restricted data or limited training resources, where fast convergence can significantly reduce computational costs. While linear regression converges slowly and is susceptible to data distribution interference, its convergence value is always low, indicating its natural limitation in complex feature modeling; BP neural network, although it has some nonlinear fitting ability, improves slowly in the initial stage (only from 0.15 to 0.55), has higher training cost, and is sensitive to the hyper-parameters and the number of samples, with the risk of training instability.

VI. CONCLUSION

This study focuses on the fatigue life prediction problem of asphalt mixture, innovatively introduces the meta-learning method, and proposes the fatigue life prediction method of asphalt mixture based on meta-learning algorithm. The method analyzes the fatigue mechanism of asphalt mixtures, extracts the influencing factors affecting fatigue, analyzes the fatigue life prediction problem of asphalt mixtures, and combines the meta-learning algorithm to construct the fatigue life prediction model of asphalt mixtures. Through systematic experimental simulation analysis, the significant advantages of the meta-learning method in terms of prediction accuracy, error control, convergence speed and generalization ability are verified. The study shows that the meta-learning method can effectively utilize the multi-task learning mechanism, quickly adapt to different working conditions, and provide an efficient and accurate solution for the fatigue life prediction of asphalt mixtures. Future research will focus on further optimizing the meta-learning model architecture and expanding its practicality and application scenarios, with a view of playing a greater role in practical engineering.

REFERENCES

[1] Qu L , Wang Y , Riara M , Mo L, Xiao Y. Study on fatigue-healing performance and life prediction of hot recycled asphalt mixture[J].*Construction and Building Materials*, 2024, 436:14.doi:https://doi.org/10.1016/j.conbuildmat.2024.136964.

[2] Wang L , Li Q , Liu K J K .Fatigue Performance Analysis and Life Prediction of Wood Tar-Based Rejuvenated Asphalt[J].*International Journal of Pavement Research and Technology*, 2024, 17(6):1531-1542. doi:https://doi.org/10.1007/s42947-023-00315-w.

[3] Silva L S V D , Bastos J B D S ,Lucas J, Jorge Luiz O, Babadopulos L F D A L, Soares, J B. Evaluation of moisture-induced damage on the fatigue life of asphalt mixtures using failure tests of asphalt binders, interfaces, and mixtures[J].*Fatigue & Fracture of Engineering Materials & Structures*, 2024, 47(7). doi:https://doi.org/10.1111/ffe.14280.

[4] Wang G , Sun E , Zhao Y .Study on fatigue performance and rutting prediction model of steel slag asphalt mixture[J].*Canadian Journal of Civil Engineering*, 2024, 52(2):228-245. doi:https://doi.org/10.1139/cjce-2024-0064.

[5] Lv S .Fatigue Life of Pre-Cut Seam Asphalt Mixture Composite Beams: A Combined Study of Fatigue Damage Evolution and Reflective Cracking Extension[J].*Buildings*, 2024, 15. doi:https://doi.org/10.3390/buildings15010050.

[6] Krzysztof S, Baran J. Towards Data-Enabled Intelligent Road Construction: A Review of Hybrid Algorithms and Applications. *Data Engineering and Applications*. 2024, 4(2): 1–13. doi:https://doi.org/10.64972/dea.2024.v4i2.19.

[7] Xu X , Wan G , Kang F , Li S, Huang W, Li Y. Evaluation Method of Fatigue Life for Asphalt Pavement on the Steel Bridge Deck Based on the Inhomogeneous Poisson Stochastic Process[J].*Materials*, 2024, 17(4):12. doi:https://doi.org/10.3390/ma17040780.

[8] Yang H , Ouyang J , Feng L , Ou J. Fatigue-Healing Performance Evaluation of Fiber-Modified Asphalt Mixture[J].*Journal of Materials in Civil Engineering*, 2024, 36(1):9. doi:https://doi.org/10.1061/JMCEE7.MTENG-16828.

[9] Petcu C , Racanel C .Estimation of Asphalt Mixture Fatigue Behaviour Using the Two-Point Bending Test[J].*Journal of Applied Engineering Sciences*, 2024, 14(1):153-160. doi:https://doi.org/10.2478/jaes-2024-0019.

[10] Spinola J R , Melo J V S D , Manfro A L .Development of an asphalt mixture with polymer incorporation modified by industrial graphene nanoplatelets for high-performance to rutting[J].*The International Journal of Pavement Engineering*, 2024, 25(1):2337303.1-2337303.26.

[11] Peng X , Jiajun C , Aihong K , Heyu K, Xukang D, Yao Z. Performance of epoxy asphalt mixtures toughened with different fibers[J].*Experimental Technology & Management*, 2024, 41(1). doi:https://doi.org/10.16791/j.cnki.sjg.2024.01.006.

[12] Wang L , Cheng P , Zhao Q .Research on Fatigue–Healing Performance of Asphalt Mixture Based on the Semicircular Bending Test[J].*Materials*, 2023, 16(19):16. doi:https://doi.org/10.3390/ma16196382.

[13] Abdulmawjoud A A .Evaluation of Fatigue Characteristics of Reclaimed Asphalt Pavement Mixtures Through Dissipated Energy[J].*International Journal of Pavement Research and Technology*, 2023, 16(1):9. doi:https://doi.org/10.1007/s42947-021-00128-9.

[14] Zhao Y , Liu X , Yan D .Evaluation of drastic temperature change on temperature stress of asphalt mixture during microwave deicing[J].*Road Materials and Pavement Design: an International Journal*, 2024(7/9):25.

[15] Ameri M , Ebrahimzadeh Shiraz M .A Review of the Studies on the Effect of Different Additives on the Fatigue Behavior of Asphalt Mixtures[J].*Advances in Civil Engineering*, 2024, 2024. doi:https://doi.org/10.1155/2024/6695747.

[16] Harne V R , Tripathi R K , Guzzarlapudi S D .Effect of extrinsic and intrinsic parameters on fatigue performance of hot mix asphalt mixtures: a review[J].*Journal of Building Pathology and Rehabilitation*, 2024, 9(1). doi:https://doi.org/10.1007/s41024-024-00406-6.

[17] Yue M , Yue J , Wang R , Guo Y, Ma X. A Comprehensive Analysis of Fatigue and Healing Capacity of Sasobit Polymer-Modified Asphalt from Two Perspectives: Binder and FAM[J].*Journal of Materials in Civil Engineering*, 2023, 35(3):19. doi:https://doi.org/10.1061/(ASCE)MT.1943-5533.0004655.

[18] Ani O J , Shafabakhsh G A , Mirabdolazimi S M .Presenting the Self-Healing Potential Model by Induction Heating Method in Warm Asphalt Mixtures by Artificial Neural Network[J].*Journal of Advanced Research in Fluid Mechanics and Thermal Sciences*, 2023. doi:https://doi.org/10.37934/arfmts.105.2.120.

[19] Li Y , Tian G , Guo W , Wei Z, Zhou Z, Chen Z. Fatigue resistance of asphalt and asphalt mixture incorporated with high-elasticity anti-rutting additive[J].*Construction and Building Materials*, 2025, 459. doi:https://doi.org/10.1016/j.conbuildmat.2024.139730.

[20] Wei H , Xu B , Li J ,Zheng J, Liu Y, Lu R. Effect of temperature on fatigue damage evolution of asphalt mixture based on cluster analysis and acoustic emission parameters[J].*Engineering Fracture Mechanics*, 2025, 317. doi:https://doi.org/10.1016/j.engfracmech.2025.110954.

[21] Xiao H , Tian W , Liu W , Guo J, Zhang Z, Jin S. Knowledge-Driven Meta-Learning for CSI Feedback[J]. *IEEE Transactions on Wireless Communications*, 2024, 23(6):16. doi:https://doi.org/10.1109/TWC.2023.3327984.

[22] Wang Z , Liu D , Hu B , Zhu C, Luo W. An improved regression model to predict fatigue life of asphalt mixes incorporating low to moderate RAP contents[J].Construction and Building Materials, 2023, 374:13. doi:https://doi.org/10.1016/j.conbuildmat.2023.130904.

[23] Liu H , Kong D , Guo T , Jin D, Chu C, Yu H. Experimental study on creep-fatigue damage of hot central plant recycling asphalt mixture containing high RAP[J].Construction and Building Materials, 2024, 440(000):13. DOI:https://doi.org/10.1016/j.conbuildmat.2024.137348.

[24] Hong L , Yuan J , Sun L , Liu L. A model predicting mechanical properties of asphalt mixtures considering void ratio and loading conditions[J].Materials Today Communications, 2024, 39. DOI:https://doi.org/10.1016/j.mtcomm.2024.109221.

[25] Monteiro S N .Influence of the Bailey Gradation Method on the Mechanical Behavior of Asphalt Mixture Containing Steel Slag as an Alternative Aggregate[J].Buildings, 2024, 14. DOI:https://doi.org/10.3390/buildings14123942.

[26] Seitllari A .Investigating the fatigue life of asphalt pavements under different aging levels using the three-point bending cylinder (3PBC) test and mechanistic-empirical analysis[J].Materials and Structures, 2025, 58(4). DOI:https://doi.org/10.1617/s11527-025-02659-0.

[27] Dabrowski B, Walczak R. Intelligent Methods for Feature Extraction and Diagnosis of Rolling Bearing Composite Faults Based on Machine Learning: A Review. Journal of Applied Automation Technologies, 2025.3:39–51. DOI:https://doi.org/10.64972/jaat.2025v3.4.

Quality Classification of Harumanis Mango Based on External Multi-Parameter and Machine Learning Techniques

Mohd Nazri Abu Bakar[1]*, Abu Hassan Abdullah[2], Muhamad Imran Ahmad[3]*, Norasmadi Abdul Rahim[4], Haniza Yazid[5], Wan Mohd Faizal Wan Nik[6], Shafie Omar[7], Shahrul Fazly Man@Sulaiman[8], Tan Shie Chow[9], Fahmy Rinanda Saputri[10]
Institute of Sustainable Agrotechnology, Universiti Malaysia Perlis, Malaysia [1, 3, 6, 7, 8]
Faculty of Electrical Engineering and Technology, Universiti Malaysia Perlis, Malaysia [2, 4]
Faculty of Intelligent Computing, Universiti Malaysia Perlis, Malaysia [3]
Faculty of Electronic Engineering and Technology, Universiti Malaysia Perlis, Malaysia [5, 9]
Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Indonesia [10]

***Abstract*—Grading Harumanis mangoes is traditionally done through manual visual inspection, which is subjective, inconsistent, and labor-intensive. Industry practices report only 70–80% consistency among human graders, with accuracy further declining under fatigue or high volumes. These limitations hinder uniform quality assurance, especially for export markets. To address this, an image-based, non-destructive grading system was developed, focusing on external features such as surface defect severity, ripeness index, shape uniformity, and size. A dataset of 1,018 mango samples was collected and analyzed using a machine vision system. Features were extracted through image segmentation and color–shape analysis, then classified using a Fuzzy Inference System (FIS) and Machine Learning (ML) models including SVM, MLPNN, and ANFIS. Enhanced SVM variants were also implemented to assess performance gains. Results showed strong performance across all parameters: ripeness index accuracy reached 93.5%, shape uniformity 91.6%, and size classification over 96%. The enhanced SVM+ achieved the best overall accuracy at 95.1% with the lowest error rates. The proposed system demonstrated clear improvements over manual grading and effectively classified mangoes into PREMIUM, GRADE 1, GRADE 2, and REJECT categories, supporting its potential for reliable real-world deployment.**

***Keywords*—Machine learning; image processing; quality assessment; Harumanis mango; appearance attributes**

I. Introduction

Mango is a widely cultivated tropical fruit that is highly valued for its unique flavor, aroma, and nutritional benefits. Among the various cultivars, the Harumanis mango is particularly renowned for its exceptional quality and popularity in Malaysia [1]. Harumanis mangoes are in high demand for both export and local consumption, and must therefore meet predetermined quality standards. Failure to meet these standards can result in substantial economic losses and reduced returns for vendors. Consequently, it is necessary to sort mangoes based on key grading parameters, including size, shape, maturity, defects, firmness, and nutritional content. The post-harvest grading process classifies mangoes into predefined quality categories, where effective grading enhances vendor and farmer confidence while expanding export opportunities. However, the traditional manual grading process is frequently subjective, laborious, and inconsistent, with reported grader consistency as low as 70–80% [2-3].

To overcome these challenges, researchers have increasingly explored advanced computer vision and ML techniques for automated mango quality assessment [4-6].[22]. While these studies demonstrated promising results, most were limited to individual features (e.g., color or size) or lacked extensive validation, reducing their robustness for real-world deployment.

Building on this foundation, the present study makes three key contributions. First, it proposes a multi-feature image-based grading framework for Harumanis mangoes that integrates surface defect severity, ripeness index, shape uniformity, and size—parameters rarely evaluated together in previous works. Second, the study evaluates multiple classification models, including FIS, SVM, MLPNN, and ANFIS, with enhanced SVM variants achieving an accuracy of 95.1%, outperforming earlier studies that typically reported accuracies below 90%. Third, to ensure robustness and reduce risks of overfitting, the study employs a k-fold cross-validation strategy with independent testing, while also acknowledging limitations such as lighting variability, dataset imbalance, and cultivar generalizability.

By addressing both methodological gaps and performance limitations in existing approaches, this study provides a more reliable, scalable, and high-accuracy grading system that offers clear improvements over traditional manual inspection and prior automated methods.

II. Related Works

Recent studies span defect synthesis, ripeness, shape, and size assessment. For surface defects, image-to-image translation with conditional GANs can augment rare defect cases [7], but synthetic artifacts risk domain shift and inflated accuracy if not validated on truly independent, real-world images. CNN-based defect classifiers [8-9] often report high accuracy on single-cultivar, lab-lighting datasets, with limited discussion of class

*Corresponding authors.

imbalance, inter-rater agreement for labels, or external test sites [21]. Classical pipelines using color/texture + SVM/kNN [10]. The study in [13] are interpretable but may be sensitive to illumination and background variability; they also evaluate a narrow set of varieties and typically do not integrate multiple attributes into a single grading decision.

For ripeness, hyperspectral + deep models [8] achieve fine spectral discrimination, yet sensor cost, throughput, and calibration burden constrain deployment; generalization across orchards and seasons is seldom shown. RGB image approaches [11-12] are cheaper but frequently rely on controlled setups and lack cross-cultivar validation. Shape-based grading [14-15] demonstrates that geometry helps detect deformities and even distinguish varieties. However, many methods assume clean silhouettes; occlusions, stem/calyx variability, and pose changes can degrade performance. Fourier descriptors [16] are transformation-invariant but may miss localized defects and can be sensitive to imperfect segmentation.

Size estimation [17-19] via dimension/mass prediction is useful, yet several works depend on explicit calibration objects, single-view assumptions, or manual placement, limiting throughput and robustness on conveyors. Across these threads, common gaps persist: (i) single-attribute focus rather than multi-attribute fusion aligned to market grades; (ii) limited head-to-head baselines vs human graders (e.g., inter-grader consistency); (iii) under-specified validation (few use stratified k-fold plus an independent test set); (iv) sparse reporting of deployment factors (speed, cost, lighting variation).

Positioning of this study. To address these gaps, this work integrates defect severity, ripeness index, shape uniformity, and size into a unified, non-destructive grading pipeline tailored for Harumanis mangoes [20]. The system is benchmarked against human grading consistency and validated through k-fold cross-validation with an independent test split, reducing overfitting risk and strengthening evidence of real-world applicability.

III. Materials and Methods

A. A Structure of Quality Grading System

This study was conducted in multiple orchards across Perlis, where the grading system hardware, shown in Fig. 1, was deployed. To acquire image data and evaluate the geometric parameters of mangoes, the system integrates both image capture and analytical models. The sorting setup consists of two main components. The first is the image processing module, which applies conventional algorithms to transform unstructured image data into structured descriptors and extract external features such as size, color index, shape similarity, and surface defects. Although external features can also be obtained using ML or deep learning methods [23-25], this study employs image processing because of its efficiency, lower computational cost, and ability to operate without extensive training datasets. Moreover, it offers greater flexibility in parameter adjustment compared to ML and deep learning approaches.

The second component combines the extracted features into a comprehensive dataset, which is subsequently used to train ML models. These models classify Harumanis mangoes into four quality categories: PREMIUM, GRADE 1, GRADE 2, and REJECT.

B. Data Preparation for Grading Process

In this study, 1018 samples of mango fruits were used. These mangoes of the Harumanis cultivar were provided by Federal Agricultural Marketing Authority (FAMA) and randomly collected from different commercial orchards located in Perlis. All the fruits without any damage and major defects were selected and classified manually into four ripeness levels by the farmer based on their visual assessment. The farmer manually classified the mangoes into four ripening levels based on peel color. All the data were recorded and classified under the ripeness levels of M1, M2, M3, and M4, respectively. Then, the sample was transported to the lab and inspected by expert workers from FAMA for quality validation.

All the samples were manually measured with respect to their length and width using a mechanical vernier caliper with an accuracy of 0.1mm. After the data on the geometry dimension was collected, all the samples were weighted using a digital scale with an accuracy of 0.1g to measure the mangoes. To predict the quality of Harumanis mangoes, a data set was made to classify the ripening level, surface defect, shape uniformity, and size that were obtained from selected mangoes. Simultaneously, the classification of quality and grading of the same samples are done by the proposed system using image processing method. This system was configured as shown in Fig. 1.

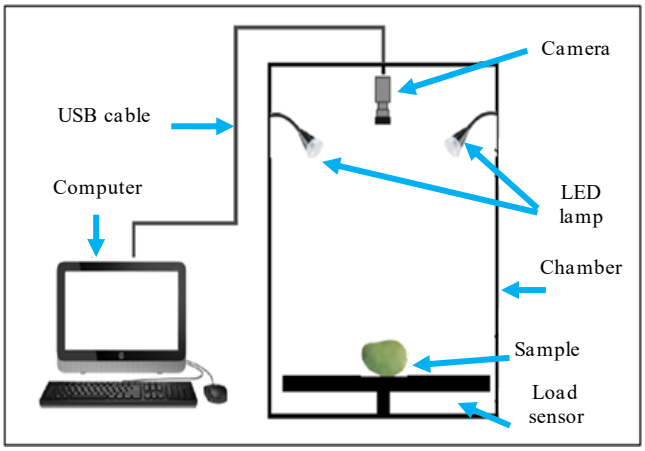


Fig. 1. System configuration for data collection.

The digital camera used for color image acquisition is the Basler acA1600-60gc GigE, with the e2v EV76C570 CMOS sensor that captures images of 1600x1200-pixel resolution. This camera is mounted 400mm on top of the chamber, and the illumination system provides two 10-watt LEDs. A generalized block diagram of the proposed research methodology for quality assessment of Harumanis is shown in Fig. 2.

C. Image Processing

In this work, the research methodology is structured as a sequential pipeline for the quality assessment of Harumanis mangoes, where each stage improves its respective task and provides input to the next. The generalized methodology (Fig. 3) consists of four main parts: 1) image acquisition, 2) image preprocessing and segmentation, 3) feature extraction, and 4) classification.

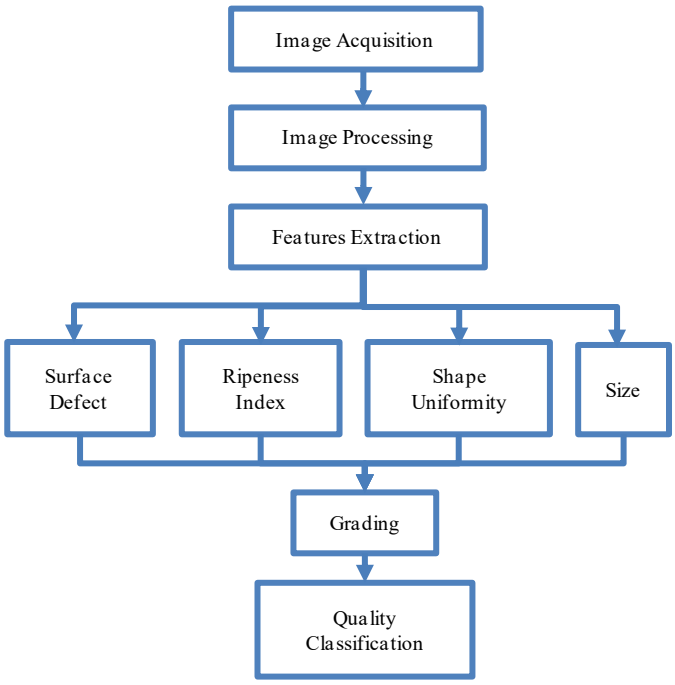


Fig. 2. Block diagram of the image processing technique for quality assessment.

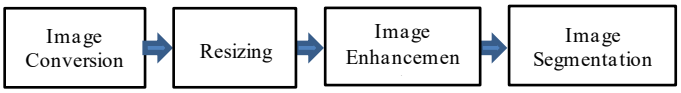


Fig. 3. Block diagram image processing technique.

In the first stage, mango images were captured using a controlled imaging setup with uniform illumination, neutral background, and fixed camera distance to minimize variation. Preprocessing included noise reduction using a Gaussian filter, contrast enhancement with histogram equalization, and background removal through thresholding to isolate the fruit. Segmentation was performed using a combination of color-based thresholding and morphological operations to accurately delineate the mango boundary, as illustrated in Fig. 4.

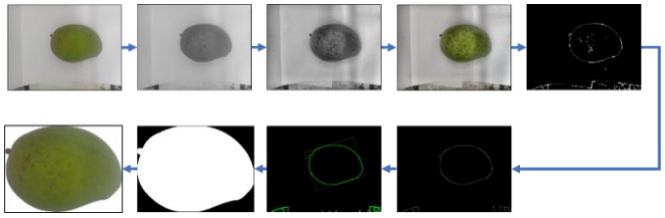


Fig. 4. Image segmentation process.

The second stage, feature extraction, involved computing color features (RGB, HSV, and statistical color moments) and shape descriptors (area, perimeter, eccentricity, Fourier descriptors) from the segmented region. These features were sampled across 1,018 mango images to ensure representativeness of different ripeness stages, sizes, and surface conditions.

Finally, in the classification stage, the extracted features were fed into a Fuzzy Inference System (FIS) to grade ripeness, shape uniformity, and defect severity, while size grading was conducted by categorizing fruit dimensions into four predefined classes. The configuration of membership functions and rule bases in FIS was optimized through iterative tuning and cross-validation to achieve robust performance.

D. Extracting Mango External Features

The feature extraction stage represents the most critical component of computer vision–based grading, wherein salient numerical descriptors are systematically derived from mango images and extraneous information is eliminated to ensure analytical relevance. In this study, two categories of features were selected—color and shape—because they directly reflect the criteria used in commercial grading of Harumanis mangoes as shown in Fig. 5. Color features were used to capture two important quality aspects: the ripeness index and external defects. Ripeness was quantified by analyzing pixel intensity distributions in RGB and HSV spaces, since skin color transitions from green to yellow or orange are strongly correlated with physiological maturity.

External defects such as bruises, dark spots, and blemishes were detected as localized deviations in color patterns, providing an objective means of assessing defect severity, which is critical for maintaining export quality standards [18], [26]. Shape features, on the other hand, were extracted to evaluate both shape uniformity and fruit size. Geometric descriptors such as aspect ratio, eccentricity, and contour smoothness were applied to determine how well the fruit conforms to the typical elongated-oval profile of Harumanis mangoes, as misshapen fruits are usually downgraded in grading practice.

Size was estimated using dimensional attributes including length, width, and projected area, since fruit size is one of the main determinants of market grade categories such as PREMIUM, GRADE 1, GRADE 2, and REJECT. By integrating these color- and shape-based features, the system provides a robust representation of the visual attributes most relevant to consumer acceptance and commercial value.

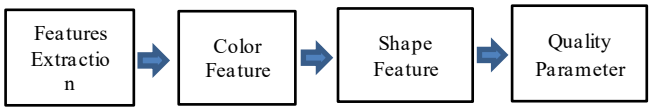


Fig. 5. Block diagram of object detection.

E. Determination of Quality Parameters

1) Surface defects: The evaluation of the Surface Defect Severity Index (SDSI) for fruit entails measuring and quantifying the skin imperfections on fruits. This is a common procedure in the fruit business to assess the quality and marketability of fruits. A numerical scoring system or grading scale is established to assess the severity of problems. This scale ranges from 0 to 5, where 0 indicates no problems, 1 signifies very minor superficial faults, 2 denotes slight superficial defects, 3 represents moderate defects, 4 indicates severe defects, and 5 reflects extremely severe defects, as seen in Table I. To evaluate this score scale, the binary picture is analyzed to determine the quantity of lesions, the cumulative size of the lesions, and the overall area of the mango. The total area of the mango and the area of the defective areas are utilized

to compute the percentage of defect, referred to as PoD. The PoD is often computed using the following equation:

$$\text{PoD} = \frac{\text{total area of defected portions}}{\text{area of the entire mango}} * 100 \quad (1)$$

Finally, the value of PoD is used as the input to the FIS to compute the membership functions for the prediction of the Surface Defect Severity Index (SDSI). The measurement of SDSI is based on three methods: PoD_{pc}, which counts pixels to measure area; PoD_{gt}, which uses mathematical analysis to understand spatial relationships; and PoD_{cl}, which divides the area into grid cells for systematic assessment. Additionally, the hybridization method (PoD_{hy}) integrates these approaches to enhance overall accuracy. Together, these methods provide a comprehensive and reliable framework for classifying SDSI.

TABLE I. CLASSIFICATION OF SDSI

Quality Class		Scoring Scale	% of Defect (*PoD*)	Inference
Commercial standard	PREMIUM	1≤	0 - 5	Free or very slight superficial defects
	1	2	5-Oct	Slight superficial defects
	2	3	Oct-15	Moderate defect
Reject	3	4	15 - 40	Severe defect
	4	5	>40	Very severe defect

2) Ripeness index: A fuzzy classifier is used to create the classification system for the ripeness level of the Harumanis mango. The fuzzy logic system has three inputs with a membership function for each input. These three inputs are represented by color measurement of *U*, *Cr* and *a** channels that were extracted from the color space of *YUV*, *YCrCb* and *L*a*b*, respectively. The determination of the ripeness index is based on the closeness of the values between the ripeness classes in terms of parameters and consumer preferences. Table II presents the mean pixel values for the *Cr, U* and *a** color channels, which severe as indicators of the ripeness level in Harumanis mangoes. There are 18 rules to be the basis for the input pattern, and the number for the output is four: unripe, underripe, ripe, and overripe. On the other hand, color channels with distinct points of range were employed with triangular membership functions. When values for the color channel are too close together, triangle membership functions have a tendency to overlap.

Concurrently, the output value is being defuzzified using the centroid approach. The set of rules that form the basis of the input pattern that produces an output was displayed by the rule viewer. The membership function combination utilizing the OR operation and the associated output are represented as the rule's viewer. It is composed of four membership functions: unripe, underripe, ripe, and overripe. Since a part of the unripe and overripe membership functions have instantly distinguishable color characteristics, they were trapezoidal in shape. Both the ripe and underripe classifications often have similar values and sometimes even the same color. They were represented by a triangular membership function.

At each level of classification, two types of data are separated into a testing set and a training set in order to validate the model. The value of training data will be a benchmark whose value is used as a differentiator for each class. To determine whether or not the system is operating correctly, test samples are created using the test data from the defuzzification process. Each of the training and test sets of data has 250 of the Harumanis mangos.

TABLE II. HARUMANIS RIPENESS INDEX ESTABLISHED VISUALLY BY THE COLOUR CHANNEL

Ripeness Index	Color Channel (Pixel)			Scoring Scale
	Cr	*U*	*a**	
M1	≤120	≤116	≤125	1
M2	>120 - 125	>116 - 120	>125 - 130	2
M3	>125 - 130	>120 - 124	>130 - 135	3
M4	>130	>124	>35	4

3) Shape uniformity: The Shape Uniformity Index (SUI) is classified as abnormal if it measures less than 80%, which may be caused by a misshapen, another clone, or grafting with another variety of mangoes. The shape uniformity is based on calculating the degree of uniformity of the sets between the Euclidean distance of the boundaries, and the differences in the parameters of the vertices of areal spatial objects. Fourier descriptors are features that can represent boundary shape of an object. They are scale, translation and rotation invariant. Contour detection was applied to the binary image to identify the largest contour (the mango), and the centroid was obtained from contour image moments. Distance of centroid from contour boundary points were computed and discrete Fourier Transform was calculated. Six standard samples selected by expert are used for comparison. These samples are labeled as T1-T6, with masses ranging from 240 g to 580 g. The Euclidean distance d_E between two vectors is defined using Eq. (2).

$$d_E(G_1, G_2) = \sum_{m=1}^{m=M_p} \sqrt{(|G_1(m)| - |G_2(m)|)^2} \quad (2)$$

where, M_p denotes the number of Fourier Descriptors pairs used for matching. G_1 and G_2 is Fourier shape descriptors, and *m* is the pixels numbered from *0* to *m – 1* at regularly spaced position. Then, the Euclidean distance, d_E from Eq. (2) is applied to measure the uniformity index of shape. The percentage of uniformity in shape *S* can be calculated as in the equation below.

$$S = (1 - d_E) * 100 \quad (3)$$

In some cases, the shape of the mango does not look like the original even though it is actually from a Harumanis clone due to the grafting effect. According to Jabatan Pertanian Perlis, they still consider as Well-shaped (WS), although it does not follow the standard shape. Therefore, the half-shape measurement of the uniformity index needs to be performed. This process is repeated with new templates consisting only of half-cut Harumanis shapes for both the lower and upper parts. Then, from Eq. (3), the percentage of uniformity in shape for entire

mango, upper and lower half-cut are calculated and defined S_{Ef}, S_{Eu} and S_{El}, respectively.

To predict SSI, fuzzy classifier is used. The values of these three parameters S_{Ef}, S_{Eu} and S_{El} for are passed as input to fuzzy system and decision has been taken that the mango fruit is *HS*, *WS* (well-shaped), and *AS* (deformed). In designing fuzzy rules for the system, considerations and assumptions obtained from [FAMA] had been taken. The system comprises of eight rules obtained from three inputs and one output of membership functions. Table III shows summary of rules represents by linguistic variable used in system.

TABLE III. IF-THEN RULES FOR CONVERTING FUZZY INPUTS TO FUZZY OUTPUTS

Input Variables (%)			Scoring scale (SUI)
S_{Ef}	S_{Eu}	S_{El}	
>80	>50	>50	1
>80	>50	<50	1
>80	<50	>50	1
>80	<50	<50	2
<80	>50	>50	2
<80	<50	>50	3
<80	>50	<50	3
<80	<50	<50	4

4) Size: Optical size grading is a process used in various industries, including agriculture, manufacturing, and quality control, to sort or classify objects based on their size using optical technology [27], [17]. This method involves the use of cameras to capture images of the objects and analyse them into different size categories. Optical size grading is preferred over manual grading in many applications because it offers a fast and efficient way to automate the sorting process, improving productivity and reducing human error. To classify mango size, a fuzzy logic classifier is used. In order to create a modelling system, fuzzy logic allows for the use of size features for variables and imprecise relationships. Two parameters, including weight and area, are employed to determine size features.

TABLE IV. RANGE OF WEIGHT AND AREA PARAMETER FOR SIZE FEATURE

Size	Weight (g)	Area (pixels)	Scoring scale (SI)
Small (S)	<300	<78212	4
Medium (M)	>301-400	>78213 - 101102	3
Large (L)	>401-500	>101103 - 127363	2
Extra Large (XL)	>501	>127364	1

The size feature is initially derived from the range of each individual parameter. These range values will serve as a reference and a range input for the fuzzy classification system. A total of 566 mangoes are used in determining the range value, weight, and area of each category, as shown in Table IV. The trapezoidal membership function of weight and area is then formed, assigning the proper range to the respective size feature.

F. Machine Learning for Grading Mangoes

The final stage focuses on classifying Harumanis mangoes into quality grades based on various features, following local standards. A dataset of 1018 samples was used to build and validate the classification models. Three supervised ML methods were applied: Multi-Layer Perceptron Neural Networks (MLPNN), Adaptive Neuro-Fuzzy Inference System (ANFIS), and Support Vector Machines (SVM). These methods were chosen for their individual strengths MLPNN handles complex non-linear patterns well [31], ANFIS integrates fuzzy logic for dealing with uncertainty and provides interpretable results [29], while SVM excels in high-dimensional data and resists overfitting with flexible kernel functions [30]. Each method classifies input features into one of several distinct quality classes. Since all dataset parameters are on similar scales, normalization may not be necessary [28].

IV. RESULTS AND DISCUSSION

The study was conducted using a dataset comprising 1,018 Harumanis mango samples. These samples were provided by the Federal Agricultural Marketing Authority and randomly collected from various commercial orchards in Perlis. All fruits selected were free from damage and major defects, and were manually classified into four ripeness indexes by the farmer based on visual assessment. The data was recorded and classified under the ripeness levels of M1, M2, M3, and M4. The results obtained for surface defect estimation, ripeness index, size, shape uniformity index classification, and final grading are discussed in the following sections.

A. Surface Defect Prediction

Fig. 6 shows the comparison of the SDSI index for the actual values (expert-provided) and the predictor methods (PoD_{pc}, PoD_{gt}, PoD_{cl}, and PoD_{hy}) is indicative of distinctions that occur in predictive accuracy across different SDSI categories. In this case SDSI 1, the actual value is around 500, and slightly overestimates PoD_{pc}, as for PoD_{gt} and PoD_{cl} they provide quite equal estimates. PoD_{hy} gives pretty much underestimates value. SDSI 2 reveals a true value much lower (~180) with PoD_{pc} and PoD_{gt} being somehow lower than the real value, whereas PoD_{cl} significantly overestimates it, followed by PoD_{hy} and which is also above the actual value.

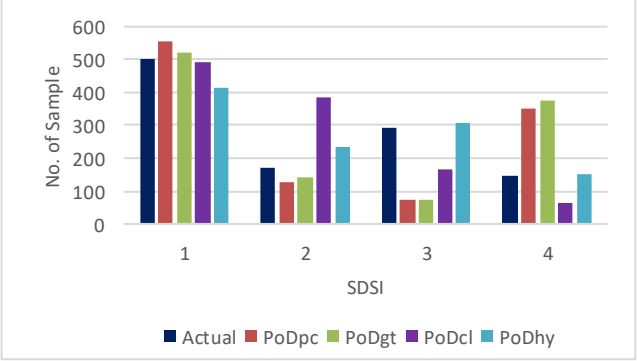


Fig. 6. Comparison of classifying the SDSI.

The actual value and the different predictor methods show SDSI 3, where most methods present lower sums than the actual value (~300), PoD_{pc} and PoD_{gt} being the farthest off. PoD_{hy} gets the nearest value to the actual one that is why it is the most accurate predictor in this category. In the last group named SDSI 4, there is a partial match for PoD_{gt} as the actual value is heavily overestimated by it. For PoD_{pc} the actual value is overestimated as well. PoD_{cl} has underestimated the value massively. PoD_{hy} is slightly overestimated but is closer than the others. In general, PoD_{hy} is found to be a very well-balanced approach that delivers predictions falling inside the correct range of values for various SDSI types. PoD_{pc} and PoD_{gt} mainly vary between the two extremes of under- and overestimation regarding the level of SDSI. PoD_{cl} is not in a steady situation, in SDSI 2 and SDSI 4 especially, it is able to either drastically underestimate or overestimate the issue.

B. Ripeness Index Prediction

From the confusion matrix in Table V, it shows that overall, the ripeness level M1-M4 achieved accuracies above than 90%. M4 groups has only three misclassification and thus attained 95.8% of accuracy. Subsequently, M3, M2 and M1 achieved accuracies with 91.7%, 92.2% and 94.4% accordingly. Seven samples from M3 were misclassified into M4 and M2, respectively. On top of that, 71 out of 77 samples from M2 are correctly classified to their group whilst two sample is misclassified into M3 and four samples in M1. Meanwhile, 51 out 54 image samples from M1 were correctly classified to their group whilst three sample was falsely classified to M2.

TABLE V. CONFUSION MATRIX FOR DETERMINING THE ACCURACY OF FUZZY

Actual Ripeness Level based on Brix	No. of Sample	Predicted Ripeness Level				Accuracy (%)
		M1	M2	M3	M4	
M1	54	51	3	0	0	94.4
M2	77	4	71	2	0	92.2
M3	84	0	3	77	4	91.7
M4	72	0	0	3	69	95.8
TOTAL	287			**Average**		93.5

C. Shape Uniformity Index (SUI) Prediction

The experiment was conducted to evaluate the effectiveness of a proposed method in distinguishing the standard shape of Harumanis mangoes from other categories, specifically well-shaped, and abnormal shape. Each category comprised 1018 samples of mangoes that varied in size. Initially, these samples underwent visual inspection by FAMA staff to assess their shape characteristics, aligning with the conditions for the SUI. The confusion matrix in Table VI showcases a discriminant model utilizing a fuzzy algorithm to predict the quality of Harumanis mangoes based on the SUI, achieving an impressive overall average accuracy of 91.6%.

The model effectively classifies samples across four quality categories: Class 1 (232 samples) with 92.7% accuracy, Class 2 (622 samples) at 87.9%, Class 3 (131 samples) with 88.5%, and Class 4 (368 samples) attaining the highest accuracy of 97.1%. Despite the model's robustness, slight misclassifications between adjacent quality classes indicate potential for improvement in accurately handling borderline cases. Overall, the results affirm the model's effectiveness in enhancing quality control process~~es~~ for Harumanis mangoes.

TABLE VI. CONFUSION MATRIX OF DISCRIMINANT MODEL FOR PREDICTING THE QUALITY OF HARUMANIS BASED ON SUI

Subjective classification	No. of Sample	Fuzzy algorithm output				Accuracy (%)
		1	2	3	4	
1	250	232	18	0	0	92.8
2	708	56	622	30	0	87.9
3	148	0	10	131	7	88.5
4	379	0	0	11	368	97.1
Average						**91.6**

D. Size Estimation

The estimated mass, M_{Est}, was further used in classifying Harumanis mangoes into four categories of size classes. Classification results from 1018 Harumanis mangoes samples show the effectiveness of the proposed size prediction using fuzzy logic in categorizing mangoes into four size classes: Small (S), Medium (M), Large (L), and Extra-Large (XL) as shown confusion matrix in Table VII. The confusion matrix reveals that the Harumanis mangoes size prediction model performs well, with most predictions being correct, especially for the XL category. The diagonal elements, which show correct predictions, are dominant, indicating high accuracy. However, some misclassifications occur, particularly between Medium (M) and Large (L) sizes, where Medium (M) mangoes are sometimes mistaken for Large (L) ones (0.99%). There are also a few cases where Large (L) mangoes are predicted as Extra-Large (XL) with 2.01%. Although there are some misclassifications, the overall accuracy is still strong. These errors are relatively small compared to the correct predictions, but they suggest that the model could benefit from further refinement, especially in distinguishing between Medium (M) and Large (L) sizes.

TABLE VII. CONFUSION MATRIX FOR SIZE PREDICTION

Actual Size		Estimated Size				Accuracy (%)
Category	**Total**	**S**	**M**	**L**	**XL**	
S	196	191	5	0	0	97.5
M	505	8	492	5	0	97.4
L	547	0	10	526	11	96.3
XL	210	0	0	3	207	98.6

TABLE VIII. CLASSIFICATION RESULTS OF COMPLETE SAMPLES INTO SIZE

Actual Size	Prediction Performance (%)		
	Precision	Recall	F1-Score
S	97.76	97.32	97.54
M	96.3	97.41	96.85
L	97.2	96.19	96.7
XL	98.93	98.23	98.39

Table VIII shows the analysis of the mango size prediction model demonstrates strong performance across all size

categories. The model exhibits high Precision, with Extra-Large (XL) mangoes having the highest value at 98.6%, indicating that when the model predicts Extra-Large (XL), it is almost always correct. Small (S) and Large (L) categories also show very high Precision scores of 97.76% and 97.20%, respectively, while Medium (M) has a slightly lower Precision at 96.30%.

In terms of Recall, which measures how well the model identifies all instances of each size, XL mangoes again perform the best with a Recall of 98.23%, ensuring that few XL mangoes are missed. The Small (S) category follows closely with a Recall of 97.42%. However, Large (L) has a slightly lower Recall of 96.19%, likely due to some misclassifications with Medium (M) mangoes. The F1-Score, which balances both Precision and Recall, also shows that the model performs well overall, with Extra-Large (XL) leading at 98.93%, followed by Small (S) at 97.54%. While Medium (M) and Large (L) have slightly lower F1-Scores of 96.85% and 96.70%, respectively, they still reflect strong performance.

E. Quality Classification

Table IX summarizes the evaluation of multiple classification models. The performance of seven different models, including ANFIS, MLPNN, SVM, SVM+ PoD_{pc}, SVM+PoD_{gt}, SVM+PoD_{cl}, and SVM+PoD_{hy}, was analyzed based on various evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Accuracy, Precision, Recall, and F1-Score. The results indicate significant differences in how well each model handles classification tasks.

TABLE IX. PERFORMANCE COMPARISON ACROSS PREDICTORS

Predictor	MAE	MSE	RMSE	Accuracy (%)	Precision	Recall	F1-Score
ANFIS	120.3	150.5	12.3	81.6	0.78	0.81	0.79
MLPNN	90.2	110.4	10.5	86.8	0.83	0.86	0.84
SVM	85.5	102.2	10.1	87.5	0.85	0.88	0.86
SVM+PoD_{pc}	98.6	115.3	10.7	83	0.81	0.83	0.82
SVM+PoD_{gt}	95.2	113.7	10.6	83.6	0.82	0.85	0.83
SVM+PoD_{cl}	180.4	210.6	14.5	65.5	0.65	0.67	0.66
SVM+PoD_{hy}	70.3	95.2	9.8	95.1	0.91	0.94	0.92

Among the base models, SVM demonstrated the highest accuracy (87.5%), outperforming both MLPNN (86.8%) and ANFIS (81.6%). SVM also had the lowest error rates, with an MAE of 85.5 and RMSE of 10.1, meaning it made fewer misclassifications compared to the other standalone models. While MLPNN performed better than ANFIS, it still had a slightly higher error rate than SVM, making SVM the best base model overall.

When additional enhancements were applied, SVM+PoD_{hy} emerged as the best model overall, achieving the highest accuracy of 95.1% and the lowest MAE (70.3) and RMSE (9.8), indicating that it had the fewest misclassifications and highest precision. Furthermore, its recall of 0.94 and F1-score of 0.92 confirm that SVM+ PoD_{hy} had the best balance between correctly identifying positives while minimizing false positives and false negatives.

This makes SVM+PoD_{hy} the most reliable predictor among all models tested. On the other hand, SVM+PoD_{cl} performed the worst, with an accuracy of only 65.5% and a significantly higher error rate (MAE of 180.4 and RMSE of 14.5). This indicates that adding the clustering-based PoD_{cl} method to SVM led to higher misclassifications, making it the least effective enhancement.

The models SVM+ PoD_{pc} (83.0%) and SVM+ PoD_{gt} (83.6%) showed moderate improvements but did not surpass the accuracy of SVM alone (87.5%), meaning that these modifications did not significantly improve performance. Their error rates were slightly lower than ANFIS but remained higher than standard SVM, suggesting that PoD_{pc} and PoD_{gt} did not contribute substantially to classification accuracy.

Overall, SVM+ PoD_{hy} was the most successful model, achieving the highest accuracy, lowest error rates, and best balance between precision and recall. In contrast, SVM+PoD_{cl} was the weakest model, suffering from high misclassification rates and poor overall performance. The findings suggest that SVM alone is a strong predictor, but hybrid methods like PoD_{hy} significantly enhance its capabilities, while others, such as PoD_{cl}, may introduce inefficiencies.

V. CONCLUSION

This study demonstrates a robust image-based framework for non-destructive quality assessment of Harumanis mangoes by integrating surface defect severity, ripeness index, shape uniformity, and size into a unified grading system. Using fuzzy inference and ML models, with SVM+ achieving the highest accuracy of 95.1%, the system clearly outperformed traditional manual grading. While the approach shows strong potential, limitations include reliance on a single cultivar and controlled imaging conditions, which may restrict generalizability. Future work will extend validation to multiple cultivars under real-world conditions, explore deep learning for richer feature extraction, and investigate real-time deployment to support scalable adoption in post-harvest supply chains.

ACKNOWLEDGMENT

The authors gratefully acknowledge the technical staff and agricultural experts from the Institute of Sustainable Agrotechnology, Universiti Malaysia Perlis, for their valuable support and contributions.

REFERENCES

[1] N. Nasron, N. S. Ghazali, N. M. Shahidin, A. A. Mohamad, S. A. Pugi, and N. M. Razi, "Soil suitability Assessment for Harumanis Mango Cultivation in UITM Arau, Perlis," IOP Conference Series Earth and Environmental Science, vol. 620, no. 1, p. 012007, Jan. 2021, doi: 10.1088/1755-1315/620/1/012007.

[2] R. Ningtyas, P. Gunawan, M. Muryeti, and S. Imam, "APPLICATION OF INTELLIGENT LABEL IN MONITORING THE PHYSICAL AND CHEMICAL QUALITY OF MANGO MANALIGA (MANGIFERA INDICA L)," Andalasian International Journal of Agriculture and Natural Sciences (AIJANS), vol. 3, no. 01, pp. 37–46, Mar. 2022, doi: 10.25077/aijans.v3.i01.37-46.2022.

[3] M. L. Ntsoane, M. Zude-Sasse, P. Mahajan, and D. Sivakumar, "Quality assesment and postharvest technology of mango: A review of its current status and future perspectives," Scientia Horticulturae, vol. 249, pp. 77–85, Jan. 2019, doi: 10.1016/j.scienta.2019.01.033.

[4] N. T. M. Long and N. T. Thinh, "Using machine learning to grade the Mango's quality based on external features captured by Vision System," Applied Sciences, vol. 10, no. 17, p. 5775, Aug. 2020, doi: 10.3390/app10175775.

[5] T. Mon and N. ZarAung, "Vision based volume estimation method for automatic mango grading system," Biosystems Engineering, vol. 198, pp. 338–349, Sep. 2020, doi: 10.1016/j.biosystemseng.2020.08.021.

[6] K. A. B. Ahmad, M. Othman, S. L. S. Abdullah, N. R. Ali, and S. R. M. Dawam, "Mango shape maturity classification using image processing," International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), pp. 1–5, Nov. 2019, doi: 10.1109/icraie47735.2019.9037776.

[7] M. H. B. Ismail, T. R. Razak, R. A. J. M. Gining, and S. S. M. Fauzi, "Simulating Bruise and Defects on Mango images using Image-to-Image Translation Generative Adversarial Networks," International Conference on Artificial Intelligence and Data Sciences (AiDAS, pp. 110–114, Sep. 2022, doi: 10.1109/aidas56890.2022.9918816.

[8] C. Yao et al., "Detection storage time of mangoes after mild bruise based on hyperspectral imaging combined with deep learning," Journal of Chemometrics, vol. 38, no. 9, May 2024, doi: 10.1002/cem.3559.

[9] R. Nithya, B. Santhi, R. Manikandan, M. Rahimi, and A. H. Gandomi, "Computer Vision system for mango fruit defect detection using deep convolutional neural network," Foods, vol. 11, no. 21, p. 3483, Nov. 2022, doi: 10.3390/foods11213483.

[10] Naik, S., & Patel, B. (2017). Thermal imaging with fuzzy classifier for maturity and size based non-destructive mango (Mangifera Indica L.) grading. In International Conference on Emerging Trends & Innovation in ICT (ICEI). https://doi.org/10.1109/etiict.2017.7977003

[11] D. Worasawate, P. Sakunasinha, and S. Chiangga, "Automatic classification of the ripeness stage of mango fruit using a machine learning approach," AgriEngineering, vol. 4, no. 1, pp. 32–47, Jan. 2022, doi: 10.3390/agriengineering4010003.

[12] F. S. Mim, S. Md. Galib, Md. F. Hasan, and S. A. Jerin, "Automatic detection of mango ripening stages – An application of information technology to botany," Scientia Horticulturae, vol. 237, pp. 156–163, Apr. 2018, doi: 10.1016/j.scienta.2018.03.057.

[13] J. Aguirre-Radilla, E. De La Cruz-Gámez, J. L. Hernández-Hernández, J. Carranza-Gómez, J. A. Montero-Valverde, and M. Martínez-Arroyo, "Texture and Color-Based Analysis to determine the quality of the Manila mango using digital image processing techniques," in Communications in computer and information science, 2022, pp. 93–106. doi: 10.1007/978-3-031-19961-5_7.

[14] M. A. Momin, M. T. Rahman, Sultana, C. Igathinathane, A. T. M. Ziauddin, and T. E. Grift, "Geometry-based mass grading of mango fruits using image processing," Information Processing in Agriculture, vol. 4, no. 2, pp. 150–160, Apr. 2017, doi: 10.1016/j.inpa.2017.03.003.

[15] N. Gururaj, V. Vinod, and K. Vijayakumar, "Deep grading of mangoes using Convolutional Neural Network and Computer Vision," Multimedia Tools and Applications, vol. 82, no. 25, pp. 39525–39550, Jul. 2022, doi: 10.1007/s11042-021-11616-2.

[16] A. D. Supekar and M. Wakode. "Multi-Parameter Based Mango Grading Using Image Processing and Machine Learning Techniques". INFOCOMP. vol. 19. no. 2. Dec. 2020.

[17] W. Aung, T. T. Thu, H. T. D. Aye, P. P. Htun and N. Aung. "Weight Estimation of Mango from Single Visible Fruit Surface using Computer Vision". Sep. 2020. 10.23919/sice48898.2020.9240422.

[18] J. D. R. Cid, W. Perez-Bailon, L. Rosas-Arias, D. R. Ocampo and J. A. Lopez-Tello. "Design of a Size Sorting Machine Based on Machine Vision for Mexican Exportation Mangoes". Nov. 2018. 10.1109/ropec.2018.8661378.

[19] K. Utai, M. Nagle, S. Hämmerle, W. Spreer, B. Mahayothee and J. Müller. "Mass estimation of mango fruits (Mangifera indica L., cv. 'Nam Dokmai') by linking image processing and artificial neural network". Elsevier BV. vol. 12. no. 1. pp. 103-110. Jan. 2019. 10.1016/j.eaef.2018.10.003.

[20] O. P. Zhen, N. Hashim and B. Maringgal. "Quality evaluation of mango using non-destructive approaches: A review". vol. 1. no. 1. pp. 1-8. Mar. 2020. 10.37865/jafe.2020.0003.

[21] B. Zheng and T. Huang. "Mango Grading System Based on Optimized Convolutional Neural Network". Hindawi Publishing Corporation. vol. 2021. pp. 1-11. Sep. 2021. 10.1155/2021/2652487.

[22] S. J. Olakanmi, D. S. Jayas and J. Paliwal. "Applications of imaging systems for the assessment of quality characteristics of bread and other baked goods: A review". Wiley. vol. 22. no. 3. pp. 1817-1838. Mar. 2023. 10.1111/1541-4337.13131.

[23] S. M. T. Islam, S. M. T. Islam, Md. Nurullah, M. Nurullah, Md. Samsuzzaman, and M. Samsuzzaman, "Mango Fruit's Maturity Status Specification Based on Machine Learning using Image Processing," 2017 IEEE Region 10 Symposium (TENSYMP), pp. 1355–1358, Jan. 2020, doi: 10.1109/tensymp50017.2020.9230951.

[24] L. Chen, T. He, Z. Li, W. Zheng, S. An and L. Zhangzhong. "Grading method for tomato multi-view shape using machine vision". Chinese Society of Agricultural Engineering. vol. 16. no. 6. pp. 184-196. Jan. 2023. 10.25165/j.ijabe.20231606.7768.

[25] M. J. Feldmann et al., "Multi-dimensional machine learning approaches for fruit shape phenotyping in strawberry," GigaScience, vol. 9, no. 5, Apr. 2020, doi: 10.1093/gigascience/giaa030.

[26] N. Puspitasari, A. M. Kiloes, and J. A. Syah, "Factors affecting sustainability of increasing mango export: an application of MICMAC method," IOP Conference Series Earth and Environmental Science, vol. 892, no. 1, p. 012101, Nov. 2021, doi: 10.1088/1755-1315/892/1/012101.

[27] N. M. Trieu and N. T. Thinh, "Quality classification of dragon fruits based on external performance using a convolutional neural network," Applied Sciences, vol. 11, no. 22, p. 10558, Nov. 2021, doi: 10.3390/app112210558.

[28] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," Applied Soft Computing, vol. 97, p. 105524, May 2019, doi: 10.1016/j.asoc.2019.105524.

[29] S. Seoni, V. Jahmunah, M. Salvi, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023)," Computers in Biology and Medicine, vol. 165, p. 107441, Sep. 2023, doi: 10.1016/j.compbiomed.2023.107441.

[30] B. Ghaddar and J. Naoum-Sawaya, "High dimensional data classification and feature selection using support vector machines," European Journal of Operational Research, vol. 265, no. 3, pp. 993–1004, Sep. 2017, doi: 10.1016/j.ejor.2017.08.040.

[31] H. Jin, Y.-G. Kim, Z. Jin, A. A. Rushchitc, and A. S. Al-Shati, "Optimization and analysis of bioenergy production using machine learning modeling: Multi-layer perceptron, Gaussian processes regression, K-nearest neighbors, and Artificial neural network models," Energy Reports, vol. 8, pp. 13979–13996, Oct. 2022, doi: 10.1016/j.egyr.2022.10.334.

Fourier Transform and Attention Guided Deep Neural Network for Face Anti-Spoofing in Medical Applications

Zhanseri Ikram
Al-Farabi Kazakh National University, Almaty, Kazakhstan
Narxoz University, Almaty, Kazakhstan

***Abstract*—Face recognition systems have become prevalent in mobile devices and security applications, increasing the demand for robust face presentation attack detection. Early efforts based on handcrafted features struggled to cope with variations in illumination, pose, and attack modalities, prompting a transition toward deep learning solutions capable of extracting subtle discriminative cues. A novel architecture built upon an EfficientNet-V2 backbone, combined with a Shuffle Attention module and Fourier heads, was developed to capture both spatial and frequency domain characteristics. A dual-path approach processes each input face image through conventional convolutional blocks and a 2D Discrete Fourier Transform path, with dedicated Fourier heads reconstructing frequency maps that reveal minute discrepancies between genuine and spoofed presentations. Experimental evaluation on the Oulu-NPU dataset demonstrates strong performance across four protocols, including robust detection under varying environmental conditions, low error rates with novel attack types, and consistent results across different sensor inputs. Metrics such as APCER, BPCER, and ACER validate the method's ability to distinguish between live and fake faces reliably. The outcomes suggest that combining spatial and frequency cues addresses limitations observed in earlier approaches, offering valuable insights for deployment in security-sensitive applications and setting a strong foundation for future research in face anti-spoofing.**

***Keywords*—Liveness detection; face anti-spoofing; deep learning; CNN; frequency domain**

I. Introduction

Face recognition systems have grown in popularity in various applications such as mobile devices and security systems. Early research in face liveness detection relied on handcrafted features such as local binary patterns and histograms of oriented gradients to capture subtle texture and motion cues [1-2], using simple image analysis and rule-based methods to separate real faces from spoof attempts. However, they often struggled with changes in illumination and pose and could not adapt well to different attack scenarios [3-4].

The field later shifted to the use of deep learning methods. Convolutional Neural Networks brought a new way to extract features from face images and proved better at detecting minute differences between live faces and fakes [5-7]. Recent studies have also focused on integrating attention mechanisms and frequency domain analysis into CNNs to improve face anti-spoofing [8-9], showing improvements in accuracy and robustness compared to traditional approaches.

In this work, EfficientNet-V2-S was used as the backbone network. EfficientNet-V2-S is known for its speed and accuracy in image processing tasks [10]. The network works by processing an input face image through several convolutional blocks that extract features at different scales. Shuffle Attention [11] introduces a novel mechanism for processing deep convolutional features by partitioning input channels and applying concurrent spatial operations. Grouping feature maps allows parallel extraction of spatial and channel information, followed by a channel shuffle that redistributes features effectively. Extensive experiments on standard networks such as ResNet-50 indicate significant performance improvements in classification and detection tasks while maintaining low computational overhead. Researchers have integrated the module into multiple vision tasks, yielding promising accuracy and efficiency results. Recent empirical evaluations validate the approach as a valuable alternative to traditional attention methods in diverse computer vision applications. Further studies corroborate findings.

Alongside the backbone and attention module, Fourier heads are added to the architecture. They reconstruct Fourier maps from the features extracted by the backbone. The Fourier maps capture frequency domain information that often holds clues to subtle differences between live faces and spoof attempts [12-14]. The Fourier heads use a series of deconvolution and dilated convolution layers to build these maps. By including frequency information, the network can detect differences that may not be clear in the spatial domain alone.

The network is built in steps that allow it to process the image, compute attention weights, and reconstruct Fourier maps. The outputs from the different stages are then combined to produce a final prediction. Such a design aims to capture both spatial and frequency cues that are important for distinguishing genuine faces from presentation attacks. Each component of the network plays a role in addressing the weaknesses of earlier methods and in improving the overall detection accuracy.

Face anti-spoofing remains a subject of great interest due to the increasing use of face recognition in security-sensitive applications. As presentation attacks grow more complex, researchers must continue to search for better ways to detect both obvious and subtle forms of spoofing. The need for robust and efficient systems is clear given the evolving methods of attack and the widespread use of biometric systems. Current study adds to this growing body of work by combining modern

network architectures, attention mechanisms, and Fourier domain analysis to address these challenges. Ongoing research in this area is essential to keep up with the rapid progress in attack techniques and to ensure the safety and reliability of face recognition systems [15].

II. Related Works

Face anti-spoofing has attracted significant research interest owing to the security implications of deploying face recognition systems. Over time, methods evolved from reliance on handcrafted features like local binary patterns and histograms of oriented gradients to modern deep learning frameworks capable of automatically extracting discriminative representations. Early techniques concentrated on identifying subtle texture and motion cues to differentiate between live and spoof presentations, though challenges with varying illumination and pose limited their practical applications.

A. Face Anti-Spoofing: Definitions and Impact

Face anti-spoofing deals with the task of telling apart live faces from fake ones produced by photographs, video replays, or masks. The goal is to protect systems that use face recognition by stopping attempts to bypass security. Research in this area has shown that proper detection is vital for preventing unauthorized access and misuse in various applications [16].

B. Traditional Methods

Earlier work in face anti-spoofing focused on hand-designed features and rule-based procedures. Texture features extracted using methods such as local binary patterns and histogram of oriented gradients were combined with classifiers like support vector machines. Thus, such systems worked by identifying subtle texture differences between real faces and printed or replayed images. While these methods were simple and fast, their performance dropped in conditions with varying lighting, pose, or image quality [17].

C. Machine Learning Methods

As the field advanced, researchers introduced machine learning methods that improved detection rates. Classifiers were built on top of handcrafted or shallow learned features extracted from facial images. Several studies combined texture information with motion cues, attempting to capture both spatial and temporal variations in the input. Fusion of multiple cues improved performance under diverse conditions, although the reliance on manually designed features still limited the overall accuracy [18].

D. Deep Learning Methods

Deep learning has introduced new possibilities in various domains, ranging from computer vision to medicine [19-21]. One of these fields involves addressing security issues such as face anti-spoofing, where models automatically learn meaningful features directly from raw data. Convolutional neural networks have been applied to capture intricate spatial details that differentiate live faces from spoof attacks. Some works introduced additional supervision or auxiliary tasks to guide the network toward discriminative patterns. Recent studies have also brought in attention mechanisms that help the network focus on regions with strong cues and methods that analyze frequency domain information to detect hidden patterns. Many deep learning approaches have reached state-of-the-art results on standard datasets such as CASIA-FASD and Replay-Attack, proving their advantage over earlier methods [22-23].

E. Challenges

Despite progress, several challenges remain. One difficulty is the wide range of attack types, from simple printed photos to complex 3D masks. Another is the gap between controlled experimental conditions and real-world scenarios. Variations in camera quality, lighting, and environmental conditions can lead to degraded performance when a model is deployed outside its training setting. Moreover, models sometimes struggle with unexpected spoofing methods not seen during training, which calls for solutions that can generalize well across different domains [24].

Another persistent challenge is the limited availability of large-scale, diverse datasets that accurately represent real-world spoofing variations. Many face anti-spoofing datasets are collected under controlled conditions, resulting in a lack of generalization when applied to heterogeneous environments such as hospitals or telemedicine platforms. Medical applications introduce unique challenges, including the presence of masks, occlusions, and patient motion artifacts, which complicate spoofing detection. Moreover, privacy constraints often limit the amount of publicly available medical facial data for model training, further exacerbating overfitting and bias issues [25-26]. To mitigate these problems, researchers have begun exploring domain adaptation, transfer learning, and synthetic data generation using generative adversarial networks (GANs) to expand dataset diversity and improve model robustness [27].

Another significant challenge is the computational complexity and real-time inference requirements for deployment in practical medical environments. Deep neural networks with attention and frequency-domain components, such as Fourier-based modules, demand high processing power, which can hinder integration into low-latency telemedicine systems or edge devices. Lightweight architectures, model pruning, and knowledge distillation are being investigated to reduce inference time while maintaining accuracy [28]. Furthermore, there is growing attention to explainability and transparency in model decisions, especially in healthcare contexts, where interpretability is critical for clinical trust and regulatory approval [29]. Future advancements will likely focus on integrating efficient model compression, interpretable attention visualization, and adaptive frequency analysis to enhance both performance and reliability in medical face anti-spoofing systems.

F. Research Gaps and Opportunities

Current work has made notable strides, yet there are several open areas for investigation. One gap lies in fully exploiting the frequency domain, as few methods integrate frequency information with spatial features for a more robust detection scheme [30]. Further studies are needed to understand how attention mechanisms can be combined with alternative representations to better capture subtle spoof cues. In addition, there is an opportunity to improve the generalization of anti-spoofing systems by addressing the domain shift between training and real-world environments [31]. Future research

should consider the integration of multiple modalities and richer representations to build systems that are both accurate and robust against evolving spoof attacks.

III. Materials and Methods

A. Proposed Model

The method illustrated in Fig. 1 begins by detecting a face using the Multitask Cascaded Convolutional Network [32]. The bounding box output by MTCNN is cropped and resized to 512 × 512 pixels. Each face image then follows two main paths. The first branch is a common path, where various stochastic augmentations (e.g., random horizontal flips, color jitter) were applied. The augmented image is forwarded through a pretrained EfficientNetV2-S backbone, which consists of eight blocks of modified convolutional layers. EfficientNetV2-S - because it offers a strong accuracy-latency trade-off for 512×512 inputs while maintaining a moderate parameter count, which is important for real-time mobile or access-control deployments. Its compound scaling and fused-MBConv layers provide large effective receptive fields and robust texture sensitivity without incurring the inference cost of deeper backbones (e.g., ResNet-101/152) or the capacity limits of lighter models (e.g., MobileNet variants). The second path is Frequency. The same 512×512 image go through a 2D Discrete Fourier Transform to produce a frequency-domain representation.

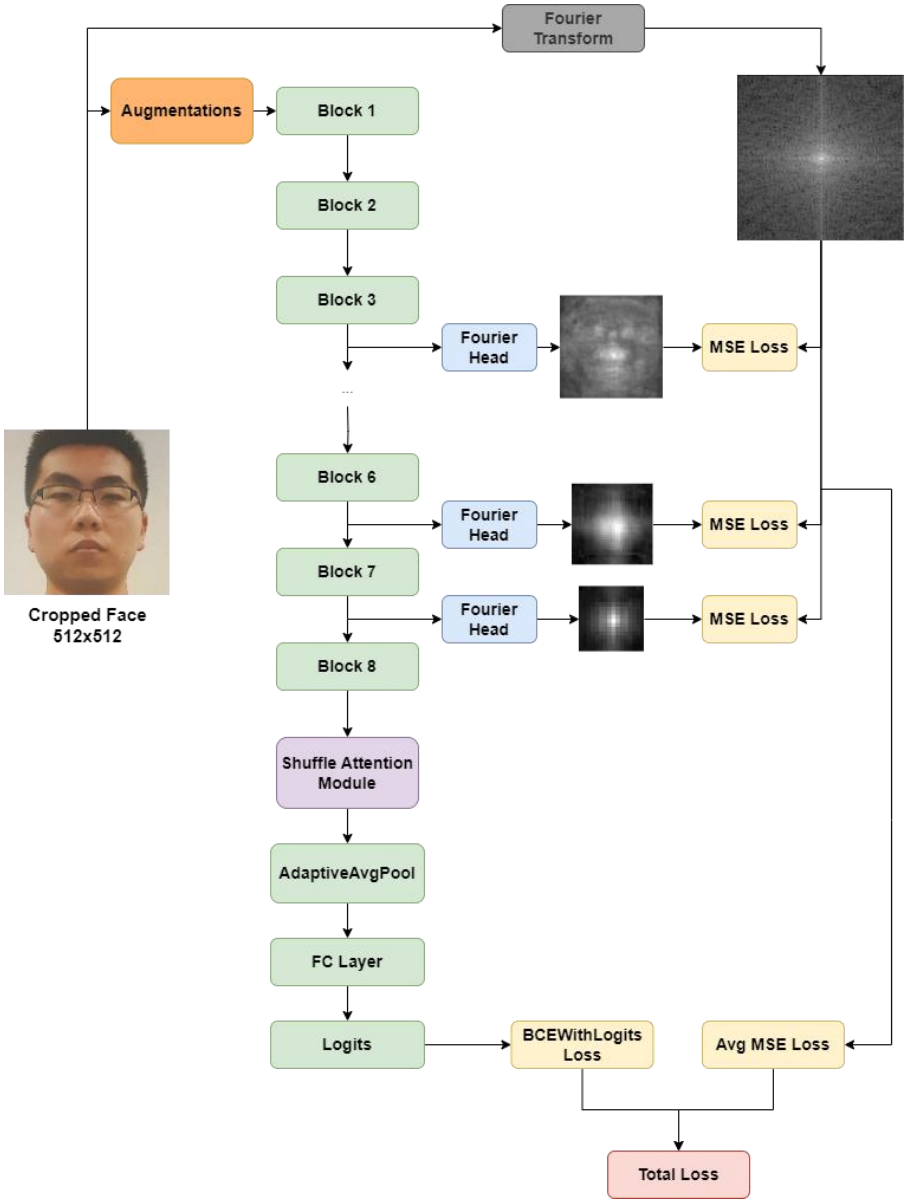


Fig. 1. Architecture of the proposed model.

After Blocks 3, 6, and 7 of the EfficientNetV2-S backbone in Fig. 2, intermediate feature maps X_3, X_6 and X_7, are each passed to a custom "Fourier Head" to reconstruct the ground-truth frequency map F. Each head contains six sequential convolutional layers (with kernel sizes of 3 × 3 or 1 × 1) interspersed with batch normalization and ReLU activation. After the third convolution, a dropout layer with probability $p = 0.5$ is inserted to reduce overfitting. The output of the fifth

convolutional layer is a single-channel feature map $\widehat{F_k} \in \mathbb{R}^{WxH}$ (where k = 3,6,7 indicates the block index), intended to approximate F. The Mean Squared Error loss for each head k is computed as Eq. (1):

$$\mathcal{L}_{MSE,k} = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}(F_{x,y} - \hat{F}_{k,x,y})^2 \quad (1)$$

After the final block, Block 8, of the EfficientNetV2-S backbone, the resulting feature map $X_8 \in \mathbb{R}^{1280x16x16}$ is processed by a Shuffle Attention module. Shuffle Attention rearranges spatial and channel dimensions for efficient information exchange without additional squeeze-and-excitation blocks. It was decided to choose Shuffle Attention over squeeze-and-excitation or CBAM because SA jointly captures channel and spatial dependencies with minimal parameter overhead and no extra bottlenecks, which is advantageous at the 16×16, 1280-channel stage of EfficientNetV2-S. The output of Shuffle Attention is a refined feature map $X_{att} \in \mathbb{R}^{1280x16x16}$.

Following the Shuffle Attention block, the refined features X_{att} pass through an Adaptive Average Pooling layer, resulting in a single vector $\boldsymbol{z}$. A fully connected layer transforms $\boldsymbol{z}$ into logits $p \in \mathbb{R}^2$ (or a single logit for binary tasks). The classification loss is the binary cross-entropy (BCE) with logits, as in Eq. (2):

$$BCELoss = -\frac{1}{M}\sum_{m=1}^{M}(y_m * \log(\sigma(p_m)) + (1 - y_m) * log(1 - \sigma(p_m))) \quad (2)$$

where, $y_m \in \{0,1\}$ is the label for sample m, p_m is the predicted logit, σ (·) is the logistic sigmoid function, and M is the number of samples in a batch. The total loss $\mathcal{L}_{total}$ combines BCE and the average MSE from the three Fourier heads, as in Eq. (3):

$$\mathcal{L}_{total} = \mathcal{L}_{BCE} + \frac{1}{3}\sum_{k\in\{3,6,7\}}\mathcal{L}_{MSE,k} \quad (3)$$

Thus, by enforcing frequency consistency, the model is encouraged to learn both spatial and frequency-based characteristics pertinent to face spoofing detection.

B. Dataset

All experiments are conducted on the Oulu-NPU [33] dataset illustrated in Fig. 3, which contains 4,950 recorded videos. From each video, 8 frames at indices {10,30,50,70,90,100,120,140} were extracted, totaling approximately 39,000 images. These videos contain:

- Six smartphone cameras (Samsung Galaxy S6 edge, HTC Desire EYE, MEIZU X5, ASUS Zenfone Selfie, Sony XPERIA C5 Ultra Dual, OPPO N3).
- Three recording sessions with varying illumination and backgrounds.
- Two display types and two printer types.
- 55 participants.
- Four standard protocols for training and testing under different environmental and cross-device conditions.

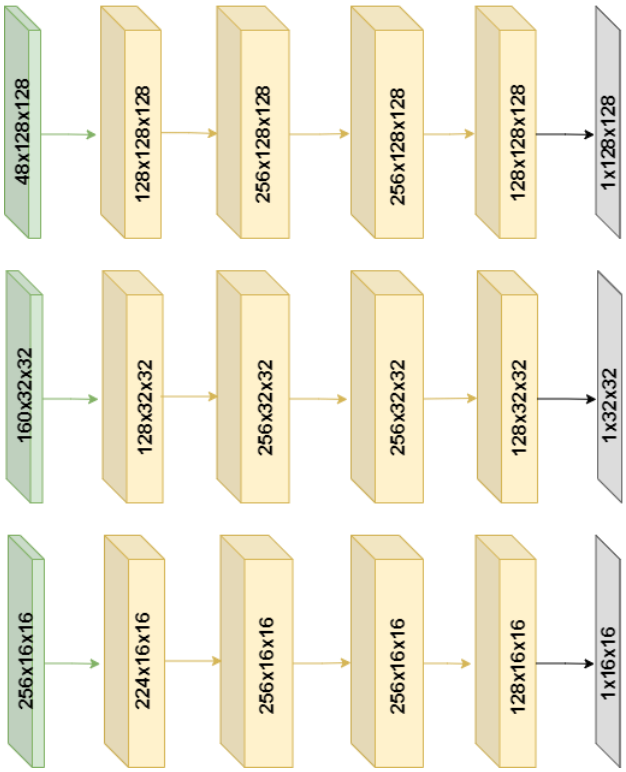


Fig. 2. Custom Fourier heads. From top to bottom: after 3rd block, after 6th block and after 7th block.

C. Experimental Setup

All models are trained and evaluated on a workstation with an NVIDIA RTX 4090 GPU (24 GB VRAM), 64 GB of CPU memory, and 24 CPU cores. The main hyperparameters are: batch size 16, Adam optimizer with weight_decay = 5×10^{-5}, number of epochs 30, ReduceLROnPlateau scheduler with patience = 3, reduce factor = 0.5. The metric used to trigger reduction in the learning rate is APCER, with an initial learning rate: 1×10^{-4}. Preliminary experiments showed that higher learning rates as 10^{-2} or 10^{-3} led to poor convergence, even when the Shuffle Attention block was removed. Consequently, a smaller learning rate with a pre-trained backbone yielded more stable results, suggesting that retaining initialization from EfficientNetV2-S is essential for this task. Each training epoch processes every sample in the dataset once. Early stopping is triggered if no improvement is observed in the validation metrics for more than five epochs, and the best model snapshot based on the validation set is preserved for final testing.

Fig. 3. Dataset samples. From left to right: live face, printed attack and replay attack.

Several transformations are applied to modify image properties while preserving essential features. Horizontal flipping reverses the image along the vertical axis, helping the model generalize to different orientations. Adjustments in contrast, gamma, and brightness alter pixel intensity distributions, simulating variations in lighting conditions. Elastic transformations, grid distortions, and optical distortions introduce non-linear deformations, making the model more resilient to geometric variations. Cutout randomly masks portions of the image, encouraging the model to focus on essential patterns rather than specific regions. Shift, scale, and rotation transformations modify the position and size of objects, aiding in generalization across different viewpoints.

IV. Results

A. Evaluation Metrics

In our evaluation of biometric authentication systems, we use three key metrics: Attack Presentation Classification Error Rate (APCER), Bona Fide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER).

APCER [Eq. (4)] measures the frequency of falsely accepting an attack attempt. A higher APCER indicates a greater likelihood of incorrectly classifying an attack as a genuine attempt. BPCER [Eq. (5)] represents the proportion of falsely rejected genuine presentations. A high BPCER means that genuine users are more frequently denied access. ACER [Eq. (6)] provides an overall error measurement by averaging APCER and BPCER. Lower ACER values indicate better system performance in distinguishing between genuine and attack presentations.

$$APCER = \frac{Number\ of\ False\ Accepts}{Total\ Number\ of\ Attack\ Presentations} \quad (4)$$

$$BPCER = \frac{Number\ of\ False\ Rejects}{Total\ Number\ of\ Genuine\ Presentations} \quad (5)$$

$$ACER = \frac{APCER+BPCER}{2} \quad (6)$$

B. Experimental Results

1) State-of-the-art comparison: The experimental outcomes were obtained from four distinct protocols designed to evaluate face presentation attack detection under various challenging conditions, as illustrated in Table I.

Protocol I focuses on generalization across environmental variations. Data recorded in three sessions with differing illumination and background settings were partitioned into training, development, and evaluation sets. The proposed method achieved zero errors, with APCER, BPCER, and ACER all measuring 0%. An analysis of the error metrics confirms an exceptional ability to differentiate between authentic and attack samples under fluctuating lighting and background conditions.

Protocol II examines the impact of novel attack artifacts generated by previously unseen print and video-replay attacks. A new print attack and a video-replay attack were incorporated into the test set to assess vulnerability to unconventional spoofing techniques. The approach recorded an ACER of 1.6%, with APCER at 0.4% and BPCER at 2.8%. A detailed evaluation of the figures reveals a robust performance in minimizing false acceptance errors, demonstrating a strong capacity to counteract artifacts introduced by diverse display and printing sources.

Protocol III investigates sensor interoperability using a Leave One Camera Out setup. Video recordings from five smartphones contributed to training and tuning, while recordings from a sixth device provided the testing material. An ACER of $0.8 \pm 0.6\%$ was obtained, with both APCER and BPCER reflecting similarly low error margins. Consistency in performance across various camera sensors confirms a reliable level of operation among different acquisition devices. A careful review of the stability of these metrics confirms the method's applicability in scenarios involving heterogeneous imaging sensors.

Protocol IV presents the most challenging scenario by combining environmental variations, attack types, and sensor diversity. Under these complex conditions, the method achieved an ACER of $1.9 \pm 2.0\%$, outperforming competing approaches that exhibited higher error rates. Error metrics for both APCER and BPCER were noticeably reduced under the simultaneous influence of multiple factors.

2) Contribution of the modules and ablation: To prove that the combination of additional modules works, the different stages by adding modules progressively were tested. In Table II, the EfficientNet-V2-S model achieves APCER of 0.8%, BPCER of 3.8%, and ACER of 2.4%. Introducing a Fourier Heads at the 3rd, 6th, and 7th blocks results in lower error rates, reducing APCER to 0.7%, BPCER to 3.2%, and ACER to 2.1%, suggesting an improvement in attack detection.

Further refinement by integrating the Shuffle Attention module into those blocks leads to the best performance, with APCER dropping to 0.4%, BPCER to 2.8%, and ACER to 1.6%. The consistent reduction in error rates indicates that the combination of Fourier Heads and SA Module contributes to more effective feature extraction, strengthening the model's ability to distinguish genuine faces from spoofing attempts.

3) Dynamic thresholding: Dynamic thresholding is a technique used in biometric security systems to adaptively determine decision boundaries based on varying conditions. Unlike fixed thresholding, which applies a single static threshold across all samples, dynamic thresholding adjusts the classification boundary according to contextual factors such as environmental conditions, device variability, or individual user characteristics. This approach enhances the robustness of biometric systems by mitigating the impact of variations in data distributions and reducing misclassification errors. In presentation attack detection, dynamic thresholding is particularly effective in balancing security and usability, as it allows the system to refine its decision-making process based on real-time inputs.

TABLE I. EXPERIMENTAL RESULTS AND COMPARISON WITH DIFFERENT STATE-OF-THE-ART METHODS ON OULU-NPU DATASET

Protocol	Method	APCER (%)	BPCER (%)	ACER (%)
1	STASN [35]	1.2	2.5	1.9
	Auxiliary [36]	1.6	1.6	1.6
	Disentangle [37]	1.7	0.8	1.3
	STDN [38]	0.8	1.3	1.1
	CDCN [39]	0.4	1.7	1
	DC-CDN [40]	0.5	0.3	0.4
	NAS-FAS [41]	0.4	0	0.2
	PatchNet [42]	0	0	0
	TransFas [43]	0.8	0	0.4
	Ours	0	0	0
2	STASN [35]	4.2	0.3	2.2
	Auxiliary [36]	2.7	2.7	2.7
	Disentangle [37]	1.1	3.6	2.4
	STDN [38]	2.3	1.6	1.9
	CDCN [39]	1.5	1.4	1.5
	DC-CDN [40]	0.7	1.9	1.3
	NAS-FAS [41]	1.5	0.8	1.2
	PatchNet [42]	1.1	1.2	1.2
	TransFas [43]	1.5	0.5	1
	Ours	0.4	2.8	1.6
3	STASN [35]	4.7±3.9	0.9±1.2	2.8±1.6
	Auxiliary [36]	2.7±1.3	3.1±1.7	2.9±1.5
	Disentangle [37]	2.8±2.2	1.7±2.6	2.2±2.2
	STDN [38]	1.6±1.6	4.0±5.4	2.8±3.3
	CDCN [39]	2.4±1.3	2.2±2.0	2.3±1.4
	DC-CDN [40]	2.2±2.8	1.6±2.1	1.9±1.1
	NAS-FAS [41]	2.1±1.3	1.4±1.1	1.7±0.6
	PatchNet [42]	1.8±1.5	0.6±1.2	1.2±1.3
	TransFas [43]	0.6±0.7	1.1±2.5	0.9±1.1
	Ours	0.8±0.5	0.8±0.8	0.8±0.6
4	STASN [35]	6.7±10.6	8.3±8.4	7.5±4.7
	Auxiliary [36]	9.3 ± 5.6	10.4 ± 6.0	9.5 ± 6.0
	Disentangle [37]	5.4±2.9	3.3±6.0	4.4±3.0
	STDN [38]	2.3±3.6	5.2±5.4	3.8±4.2
	CDCN [39]	4.6±4.6	9.2±8.0	6.9±2.9
	DC-CDN [40]	5.4±3.3	2.5±4.2	4.0±3.1
	NAS-FAS [41]	4.2±5.3	1.7±2.6	2.9±2.8
	PatchNet [42]	2.5±3.8	3.3±3.7	2.9±3.0
	TransFas [43]	2.1±2.2	3.8±3.5	2.9±2.4
	Ours	2.5±2.5	1.2±2.1	1.9±2.0

TABLE II. INDIVIDUAL CONTRIBUTION OF EACH MODULE TESTED ON PROTOCOL II

Method	APCER (%)	BPCER (%)	ACER (%)
EfficientNet-V2-S	0.8	3.8	2.4
EfficientNet-V2-S w/ 3rd, 6th, 7th blocks Fourier Heads	0.7	3.2	2.1
EfficientNet-V2-S+SA Module w/ 3rd, 6th, 7th blocks Fourier Heads	**0.4**	**2.8**	**1.6**

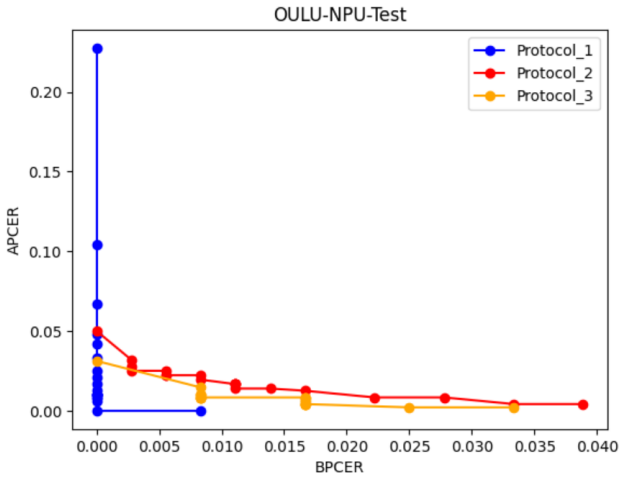


Fig. 4. Dynamic thresholding for 3 protocols.

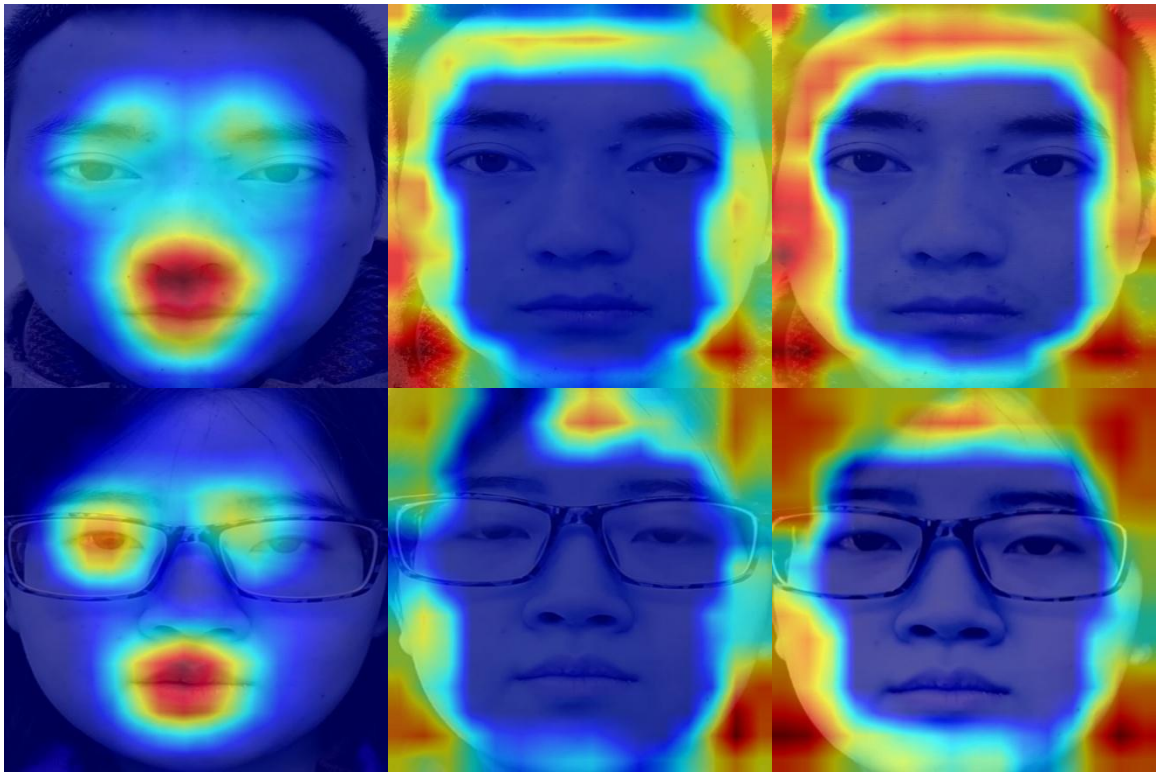
Fig. 4 illustrates the performance of three different protocols under dynamic thresholding on the OULU-NPU-Test dataset. The *x*-axis represents the BPCER, which indicates the proportion of genuine samples misclassified as attacks. The *y*-axis corresponds to the APCER, reflecting the proportion of attack presentations misclassified as genuine. Lower values on both axes indicate a more effective system, as it minimizes both types of errors.

Protocol I, represented in blue, demonstrates a steep decline in APCER as BPCER increases. At very low BPCER values, APCER remains considerably high, indicating that the system is prone to false acceptance when a strict threshold is applied. This pattern suggests that the protocol may be highly sensitive to variations in presentation attacks but suffers from instability when distinguishing bona fide samples. The rapid decrease in APCER as BPCER increases further suggests that a slight relaxation of the threshold significantly improves attack detection, albeit at the cost of increased false rejection.

Protocol II, visualized in red, follows a more stable trajectory, exhibiting a gradual reduction in APCER as BPCER increases. Compared to Protocol I, this approach achieves a more balanced trade-off between false acceptance and false rejection, indicating that it may be more reliable in practical deployment scenarios. The smoother trendline suggests that Protocol II is less sensitive to minor variations in input data, making it a more consistent option for presentation attack detection.

Protocol III, shown in orange, consistently maintains the lowest APCER across the observed BPCER range. The results indicate that this protocol outperforms the other two in minimizing attack presentation errors while maintaining a relatively low bona fide classification error. The stability of the curve suggests that Protocol III achieves a well-calibrated balance between security and accessibility, making it the most effective among the three.

4) Grad-Cam results: Grad-Cam [34] results in Fig. 5 provide an insightful interpretation of the deep learning model's decision-making process for detecting presentation attacks. Heatmaps indicate which facial regions contribute most to classification, revealing distinct patterns across live, print, and video-replay samples.



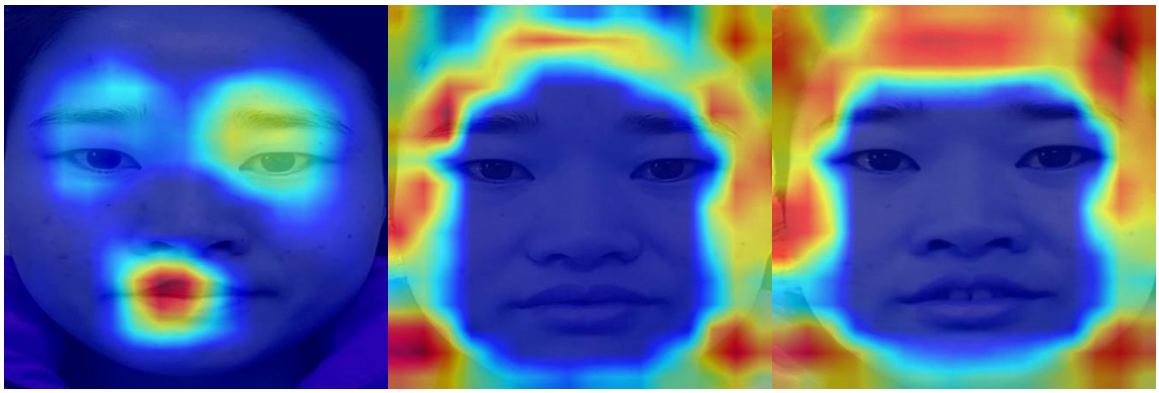


Fig. 5. Grad-Cam results. From left to right columns: live, print, and video-replay.

In live face images, attention is primarily focused around the mouth and nose, where color, texture, and depth cues play a significant role in distinguishing genuine faces. The uniform blue background suggests that the model effectively ignores non-relevant areas, reinforcing its reliance on intrinsic facial details.

For print and video-replay attacks, activations shift towards facial boundaries, where red and yellow highlights suggest the model detects inconsistencies in edges and lighting artifacts. The presence of strong activations around the periphery indicates an effort to identify unnatural textures, which are often present in spoofing attempts.

Images with glasses introduce additional complexity, with attention distributed around the frames and nose bridge. Higher activations in print and video attacks suggest reflections and distortions are key factors in classification.

Variations in heatmap intensity between attack types confirm the model's ability to differentiate genuine and spoofed faces. Attention shifts in predictable ways, aligning with expected signs of presentation attacks, demonstrating the model's capacity to detect spoofing artifacts with reliability.

Overall, the experimental results indicate several advantages of the proposed method. The approach demonstrates flawless performance in the face of environmental variations, maintains a low false acceptance rate against novel attack artifacts, and exhibits consistent error rates across different sensor inputs. When confronted with a combination of challenges, it delivers competitive performance with reduced error metrics, suggesting promising applicability for real-world face presentation attack detection systems.

V. Discussion

The experimental results indicate a promising advancement in face presentation attack detection. Findings confirm that spatial and frequency domain information combined into a unified architecture can significantly improve detection accuracy. The adoption of EfficientNet-V2-S as the backbone, combined with a Shuffle Attention module, has provided a robust framework capable of extracting discriminative features even under adverse conditions. Integration of Fourier heads to reconstruct frequency maps introduces a novel perspective that captures subtle differences between live and spoofed faces, as evidenced by the low error rates across various protocols.

Performance under controlled environmental variations demonstrated strong discrimination between genuine and attack samples. Minimal errors in protocols addressing different illumination and background conditions suggest that the network effectively adapts to changes in lighting and scene composition. Robust outcomes in scenarios involving previously unseen print and video-replay attacks further support the method's aptitude for recognizing unconventional spoof artifacts. Results reveal a capacity for mitigating false acceptance errors, which remains critical in maintaining the integrity of biometric systems.

Interoperability across different sensor types has been confirmed through experiments using diverse smartphone cameras. Consistent performance in sensor-specific testing scenarios reinforces the network's reliability, even when presented with imaging data acquired from heterogeneous sources. Outcomes recorded in the most challenging scenario, where multiple factors were simultaneously present, indicate competitive error metrics that rival or surpass existing methods. Low Average Classification Error Rates across protocols suggest that the design effectively balances the detection of genuine presentations with the rejection of spoof attacks.

Observations imply that the fusion of spatial features and frequency representations addresses limitations inherent in earlier approaches based solely on handcrafted or spatial features. The network's ability to capture minute texture and motion cues, while simultaneously processing global frequency information, contributes to its overall robustness. A careful analysis of the error metrics shows potential avenues for further exploration, particularly regarding the model's performance in real-world environments where attack types may be more diverse.

Limitations related to domain shifts between training and deployment environments persist, inviting additional research to refine generalization capabilities. Future investigations might consider the incorporation of alternative modalities or additional attention mechanisms to further improve detection rates under evolving spoof conditions. Overall, the findings present compelling evidence that integrating modern network architectures with both spatial and frequency domain analysis provides a viable pathway toward more reliable and secure face presentation attack detection systems.

VI. Conclusion

This work demonstrates that explicitly coupling spatial features with frequency-aware supervision yields robust face presentation attack detection across environmental, attack-type, and sensor variations. An EfficientNetV2-S backbone provides an effective accuracy-latency balance for deployment; Fourier heads guide the network to preserve high-frequency spoof cues that are often attenuated in spatial pipelines; and, Shuffle Attention efficiently enhances discriminative regions without heavy computation. Across Oulu-NPU's four protocols, the system attains 0% ACER in Protocol I, 1.6% in Protocol II, 0.8 ± 0.6% in Protocol III, and 1.9 ± 2.0% in Protocol IV, indicating strong generalization to novel artifacts and unseen sensors. Beyond raw numbers, the ablations and interpretability results clarify why the method works: frequency supervision reduces bona-fide rejections under photometric variation, while attention improves rejection of attack-specific structures, together producing consistent gains.

For security-sensitive biometric applications, reducing APCER directly lowers the risk of unauthorized access, and lowering BPCER preserves user experience. The proposed design advances practical anti-spoofing by improving both, without sacrificing throughput or model size, making it suitable for edge and mobile deployments. Performance under severe domain shift can still degrade. Future work will: i) incorporate temporal cues (e.g., rPPG or micro-motion) alongside frequency supervision, ii) explore self-supervised or domain-generalization objectives to handle unseen attacks, and iii) calibrate dynamic thresholds using cohort/device context to further stabilize APCER-BPCER trade-offs in the field.

Acknowledgments

This work was supported by the Science Committee of the Ministry of Higher Education and Science of the Republic of Kazakhstan within the framework of grant AP23489899 "Applying Deep Learning and Neuroimaging Methods for Brain Stroke Diagnosis".

References

[1] Patel, D., Shah, H., & Kumar, P. (2018). Liveness detection: A review of traditional and deep learning methods. Journal of Visual Communication and Image Representation, 54, 234–245.

[2] Kumar, P., Patel, D., & Shah, H. (2017). Anti-spoofing techniques for biometric systems. IEEE Transactions on Information Forensics and Security, 12(8), 1870–1884.

[3] Li, S., Liu, Z., & Zhang, H. (2019). A survey on face anti-spoofing methods. Pattern Recognition Letters, 115, 61–72.

[4] Anjos, A., Menotti, D., & Batista, G. (2020). Deep learning methods for face anti-spoofing. IEEE Transactions on Biometrics, Behavior, and Identity Science.

[5] Wang, L., Yu, G., & Li, X. (2021). Fourier analysis in face anti-spoofing. IEEE Access, 9, 102354–102363.

[6] Altayeva, A., Omarov, B., Jeong, H. C., & Im Cho, Y. (2016). Multi-step face recognition for improving face detection and recognition rate. Far East Journal of Electronics and Communications, 16(3), 471.

[7] Yang, Y., Chen, Z., & Liu, Z. (2020). Attention mechanisms in convolutional neural networks: A review. IEEE Signal Processing Magazine, 37(3), 75–85.

[8] Zhang, Y., Zhou, K., & Lin, J. (2021). A novel Fourier head for face anti-spoofing. In Proceedings of the International Conference on Pattern Recognition.

[9] Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller models and faster training. arXiv preprint arXiv:2104.00298.

[10] Zhang, X., Yang, J., & Wang, L. (2020). Shuffle attention for neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[11] Zhang, Q.-L., & Yang, Y.-B. (2021). SA-Net: Shuffle attention for deep convolutional neural networks. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2235-2239. https://doi.org/10.1109/ICASSP39728.2021.9414568

[12] Feng, W., & Xie, X. (2019). Frequency domain analysis for robust face anti-spoofing. IEEE Transactions on Information Forensics and Security, 14(10), 2802–2813.

[13] 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).

[14] Guo, Y., Zhao, G., & Zhang, R. (2020). Exploring frequency cues in face anti-spoofing with deep networks. IEEE Access, 8, 75678–75689.

[15] Cheng, W., & Li, X. (2022). Continuous advances in face anti-spoofing:A review of recent methods. IEEE Transactions on Biometrics, Behavior, and Identity Science, 4(1), 1–14.

[16] Boulkenafet, I., Komulainen, J., & Hadid, A. (2016). Face spoofing detection from single images using spatio-temporal features. IEEE Transactions on Information Forensics and Security, 11(12), 2718–2732.

[17] Chingovska, I., Anjos, A., & Marcel, S. (2012). On the effectiveness of local binary patterns in face anti-spoofing. In 2012 IEEE International Joint Conference on Biometrics (IJCB) (pp. 1–7). IEEE.

[18] Patel, V. M., & Chellappa, R. (2013). Face anti-spoofing using support vector machines with handcrafted features. In Proceedings of the IEEE Workshop on Biometrics (pp. 1–6).

[19] Al Noman, M. A., Zhai, L., Almukhtar, F. H., Rahaman, M. F., Omarov, B., 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.

[20] Akshay Dilip Lahe, Guddi Singh, "A Survey on Security Threats to Machine Learning Systems at Different Stages of its Pipeline", International Journal of Information Technology and Computer Science(IJITCS), Vol.15, No.2, pp.23-34, 2023. DOI:10.5815/ijitcs.2023.02.03

[21] 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).

[22] Atoum, Y., Liu, S., & Maybank, S. (2017). Face spoofing detection using patch-based convolutional neural networks. In 2017 International Joint Conference on Biometrics (IJCB) (pp. 1–8). IEEE.

[23] Li, J., Liao, R., & Liu, M. (2018). Deep texture features for face anti-spoofing. IEEE Transactions on Image Processing, 27(9), 4489–4501.

[24] Yang, Y., Liu, Z., & Chen, X. (2018). Challenges in face anti-spoofing: A review. IEEE Transactions on Biometrics, Behavior, and Identity Science, 1(2), 107–119.

[25] 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.

[26] Bala Dhandayuthapani V., "Enhancing Jakarta Faces Web App with AI Data-Driven Python Data Analysis and Visualization", International Journal of Information Technology and Computer Science(IJITCS), Vol.16, No.5, pp.36-51, 2024. DOI:10.5815/ijitcs.2024.05.03

[27] Lin, J. D., Lin, H. H., Dy, J., Chen, J. C., Tanveer, M., Razzak, I., & Hua, K. L. (2021). Lightweight face anti-spoofing network for telehealth applications. IEEE Journal of Biomedical and Health Informatics, 26(5), 1987-1996.

[28] Ifeoluwani Jenyo, Elizabeth A. Amusan, Justice O. Emuoyibofarhe, "A Trust Management System for the Nigerian Cyber-health Community",

International Journal of Information Technology and Computer Science(IJITCS), Vol.15, No.1, pp.9-20, 2023. DOI:10.5815/ijitcs.2023.01.02

[29] Adeniran, A. A., Onebunne, A. P., & William, P. (2024). Explainable AI (XAI) in healthcare: Enhancing trust and transparency in critical decision-making. World Journal of Advanced Research and Reviews, 23(3), 2647-2658.

[30] Liu, M., & Liu, X. (2020). Future directions in face anti-spoofing research. IEEE Signal Processing Magazine, 37(5), 60–70.

[31] Wang, Y., Zhang, Z., & Li, W. (2021). Integrating frequency and spatial cues for robust face anti-spoofing. IEEE Transactions on Multimedia, 23, 1572–1583

[32] Zhang, K., Zhang, Z., & Li, Z. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. In 2016 IEEE Signal Processing Letters (Vol. 23, pp. 1499-1503). IEEE. https://doi.org/10.1109/LSP.2016.2603342

[33] Boulkenafet, Z., Komulainen, J., Li, L., Feng, X., & Hadid, A. (2017). OULU-NPU: A mobile face presentation attack database with real-world variations. In 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017).

[34] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision, 128(4), 1187–1200. https://doi.org/10.1007/s11263-019-01228-7

[35] Yang, X., Luo, W., Bao, L., Gao, Y., Gong, D., Zheng, S., Li, Z., Liu, W.: Face anti-spoofing: Model matters, so does data. In: proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3507–3516 (2019)

[36] Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 389–398 (2018)

[37] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Ying Tai, Shouhong Ding, Jilin Li, Feiyue Huang, Haichuan Song, and Lizhuang Ma. Face anti-spoofing via disentangled representation learning. In ECCV, 2020. 1, 5, 6

[38] Y. Liu, J. Stehouwer, and X. Liu, “On disentangling spoof trace for generic face anti-spoofing,” in Proc. ECCV, 2020, pp. 406–422.

[39] Z. Yu et al., “Searching central difference convolutional networks for face anti-spoofing,” in Proc. CVPR, 2020, pp. 5295–5305.

[40] Z. Yu, Y. Qin, H. Zhao, X. Li, and G. Zhao, “Dual-cross central difference network for face anti-spoofing,” in Proc. IJCAI, 2021, pp. 1281–1287.

[41] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li, and G. Zhao, “NAS-FAS: Static-dynamic central difference network search for face anti-spoofing,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 9, pp. 3005–3023, Sep. 2021.

[42] C.-Y. Wang, Y.-D. Lu, S.-T. Yang, and S.-H. Lai, “PatchNet: A simple face anti-spoofing framework via fine-grained patch recognition,” in Proc. CVPR, 2022, pp. 20281–20290.

[43] Z. Wang, Q. Wang, W. Deng, and G. Guo, "Face Anti-Spoofing Using Transformers With Relation-Aware Mechanism," IEEE Transactions on Biometrics, Behavior, and Identity Science, vol. 4, no. 3, July 2022, p. 439.

A Technique for Automated Parallel Optimization of Function Calls in C++ Code

Shuruq Abed Alsaedi[1], Fathy Elbouraey Eassa[2], Amal Abdullah AlMansour[3],
Lama Abdulaziz Al Khuzayem[4], Rsha Talal Mirza[5]
Department of Computer Science-Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah, Saudi Arabia[1, 2, 3, 4, 5]
Department of Computer Science-College of Computer Science and Engineering,
Taibah University, Yanbu 46421, Saudi Arabia[1]

***Abstract*—In modern software development, achieving high performance increasingly relies on effective parallelization. While much of the existing research has focused on loop-level parallelism, function-level parallelization remains relatively underutilized. Yet, in many real-world applications, function calls serve as natural units of computation that could greatly benefit from concurrent execution. To address this gap, we present an automated tool that analyzes sequential C++ code, identifies independent function calls, and evaluates their suitability for parallel execution. The tool performs three key analyses: dependency analysis to detect function calls, context analysis to understand execution conditions, and workload assessment to determine whether parallelization would result in significant performance benefits. Based on the analysis results, the tool transforms eligible function calls into parallel equivalents without altering the original program logic. Additionally, the tool generates detailed Control Flow Graphs (CFG) for each function in three formats, facilitating further structural analysis. Three benchmark programs were used in experimental testing. The evaluation measured both sequential and parallel execution times, along with the computed performance gain expressed as a percentage reduction in runtime. Results demonstrated the tool's ability to improve execution efficiency and reduce processing time. These outcomes emphasize the tool's role in advancing function-level automatic parallelization. The tool showed notable performance improvements across the three benchmark applications, with the Employee Performance System achieving the highest improvement of 54.6%, followed by the Genomic Sequence System at 48.3%, and the Book Reviews System achieving an improvement of 36.1%. Demonstrating the tool's ability to improve efficiency via automated function-level parallelization.**

Keywords*—Automatic parallelization; function-level parallelization; C++ code optimization; parallel computing; control flow graph; dependency analysis; performance optimization

I. Introduction

Improvements of multi-core processors have increased the adoption of parallel programming among software developers. However, developing parallel code remains a challenging and error-prone task. As a result, automating code parallelization has emerged as a hot topic in high-performance applications research [1]. Available parallelization techniques usually target specific sections of code that are suitable to concurrent execution. Loops are the most common candidates. Loops often contain iterative operations that can be executed in parallel without interdependencies, making them ideal for leveraging multi-core and multi-threaded architectures.

However, there are other code constructs, such as function calls, that can enhance performance and efficiency if they are properly utilized.

By focusing on function calls, parallelization strategies can effectively improve the software applications. However, despite years of research, the automatic parallelization of function calls remains an unresolved problem. Function calls often still require manual parallelization [1].

The motivation for this work stems from several critical challenges in modern software development:

1) Underutilization of parallelization opportunities: While parallelization research has focused heavily on optimizing loops, function calls have received comparatively less attention. This highlights a significant gap in current parallelization research, as function calls provide substantial opportunities for enhancing performance. In many real-world applications, computational workload is distributed across numerous function calls rather than concentrated in loops. For instance, data processing applications often involve multiple independent operations such as data extraction, filtering, aggregation, and statistical analysis, each implemented as separate functions. These functions can potentially execute concurrently without dependencies, yet they remain sequential in most codebases due to the lack of automated transformation tools. Addressing the parallelization of function calls helps improve the efficiency and scalability of code. Nonetheless, developing parallel code or converting certain sequential code to run in parallel is a challenging task that requires significant effort from programmers.

2) Complexity of manual parallelization: The complexities of parallel programming start from the problem definition, as not all problems are suited for parallelization. Programmers must carefully consider different aspects of parallel processing, such as how to distribute the current workload across parallel threads and manage communication between these threads [2]. This manual effort is time-consuming, error-prone, and requires deep expertise in both parallel programming paradigms and the application domain. Developers must analyze data

dependencies, identify race conditions, manage synchronization, and ensure thread safety tasks that significantly increase development time and maintenance costs.

3) Performance demands of modern applications: Contemporary software applications face increasing computational demands. From big data analytics and scientific computing to machine learning and real-time systems, the need for efficient parallel execution has never been greater. C++ programming language is known for its performance and fine-grained control and become an essential component in the development of high-performance computing applications, including those that leverage parallelism [8]. However, the gap between available hardware capabilities and software utilization remains substantial. Many existing sequential C++ applications could benefit significantly from parallelization but lack the resources or expertise for manual transformation.

Our proposed tool was designed to analyze and transform C++ code, which was selected due to its relevance in performance-sensitive software and its rich set of parallelization libraries. This gap is particularly significant because:

- Natural Computation Units: Functions represent natural units of computation with well-defined interfaces and encapsulated logic, making them ideal candidates for parallel execution.
- Coarse-Grained Parallelism: Function-level parallelism offers coarse-grained task parallelism that can better utilize modern multi-core architectures with lower synchronization overhead compared to fine-grained loop iterations.
- Code Maintainability: Parallelizing at the function level preserves code structure and maintainability better than aggressive loop transformations.

This study proposes a fully automated tool that offers an advanced method for analyzing and selecting suitable function calls for parallelization in C++ code, a language with unique structural and concurrency characteristics. The proposed tool integrates static analysis, control flow graph generation, and a lightweight scoring mechanism to identify and transform independent function calls into parallel equivalents using *std::async*. By applying this technique to C++, the tool bridges a critical gap in existing research. This contribution offers practical value for developers working with performance-critical C++ applications, especially in cases where loop-level parallelism is insufficient.

The main contributions of this work are:

1) Comprehensive static analysis framework: We introduce a three-phase analysis approach combining:

- Dependency Analysis: Automatically detects function calls and analyzes their data and control dependencies to ensure correctness of parallel transformations.
- Context Analysis: Evaluates execution contexts to determine parallelization safety and identify functions that can be executed asynchronously without affecting program semantics.
- Workload Assessment: Employs a work potential metric that quantifies computational complexity, enabling prioritization of parallelization candidates.

2) Automated code transformation pipeline: The tool provides end-to-end automation from analysis to code generation:

- Automatically transforms eligible sequential function calls into parallel equivalents using C++ standard library features (std::async, std::future), without requiring manual intervention or code restructuring.
- Preserves original program logic and semantics while introducing parallelism.

3) Visual Analysis and Debugging Support:

- Generates detailed Control Flow Graphs (CFGs) for each function in three formats (DOT, PDF, and TXT), facilitating structural analysis and verification.
- Provides comprehensive reporting of parallelization decisions, including work potential scores and context classifications.
- Enables developers to understand, validate, and fine-tune the automated transformations.

In this study, three benchmark programs were selected to conduct experiments. The experiments aimed at measuring parallel execution time, sequential execution time, and computed performance improvement, representing the percentage reduction in execution time. The results show that our proposed tool successfully enhances performance and reduces execution time. The findings highlight the tool's capability as a step forward in bridging the current gap in automatic parallelization research, particularly at the function-call level.

The structure of this paper is as follows: Section II provides background information; Section III reviews related work; Section IV describes the system architecture and methodology; Section V presents the experimental results and discussion; Section VI evaluates performance improvements; and Section VII concludes the paper.

II. Background

Historically, processor design has focused on combining more advanced features, achieving higher clock speeds and increasing thermal limits. However, the pursuit of enhanced performance remains a necessary demand. This has driven the adoption of integrated multi-processor (multi-core) architectures [3].

However, transistors can't keep getting smaller indefinitely. Nowadays, as transistors become thinner, chip makers aim to manage heat generation and power usage; these are two major challenges. Additionally, performance enhancement techniques like running multiple instructions at the same time have reached their limits.

Due to these problems, the rate of processor performance improvements has started to slow down. In the 1990s, chip performance increased by 60 percent each year, while from

2000 to 2004, this growth slowed to 40 percent annually, with performance only rising by 20 percent.

In fact, improvements in computer performance have essentially resulted from making chips smaller and increasing their number of transistors. According to Moore's Law, this approach has led to faster chip speeds and reduced costs. Instead of making one super powerful core, companies are now creating chips with multiple cores that use less energy and run cooler. Even though these multicore chips might not be as fast as the top single-core ones, they improve overall performance by handling more tasks at the same time [4].

Many modern programs require more computing power than traditional single-threaded computers can offer. In order to achieve better performance, parallel processing is necessary to meet demand. Parallelism allows executing multiple tasks at the same time to solve complex problems in less time. Multicore processing requires running programs on more than one core within a single CPU chip, which is a type of parallel processing. Multicore utilization means using the CPU's multiple cores efficiently [5].

In parallel processing of programs, instructions are split across many processors to run the same program faster than with sequential processing. Some programs that require tasks such as sorting and searching are designed to use parallel processing. Such programs are often tested on large databases. Experimental results using bubble sort and linear search show that sharing the load and utilizing multicore parallel processing make tasks easier and quicker compared to sequential processing. Parallel processing and multicore processing both involve the ability to run code simultaneously on different CPUs, machines, or cores. Multicore processing is designed to perform tasks in parallel. A core can be considered as a small processor within a larger processor. Therefore, optimizing code for multicore processing is closely tied to parallelization [4] [5].

Finding opportunities to run code automatically in parallel is essential for developing efficient programs to run on many multithreaded applications. This area has been researched for many years. Even though hardware provides some fine-grained parallelism, compiler researchers have to develop techniques to automatically convert source code into suitable parallel applications.

With the growing ubiquity of distributed computing, the software industry demands compiler technologies capable of automatically parallelizing software. However, the demand for automatic parallelization in compilers is growing as clusters and other kinds of distributed computing become widely adopted, and CPU technology moves toward higher levels and coarser granular parallelism [6].

Generally, automatic parallelization involves the following three procedures [7]:

- Dependency analysis to identify potential parallelism regions: Carrying out a dependency analysis to identify regions in the code that are capable of parallelization.
- Restructuring the program into blocks that can run in parallel: Organizing the code into number of blocks that can run simultaneously. In this phase, different transformations are applied to maximize the level of parallelism achievable within the code.
- Generating parallel code suitable for specific architectures: Producing parallel code optimized for a particular architecture by assigning program blocks to available processing units and developing appropriate strategies to enable parallelism based on the targeted system.

Building on these core steps, it is essential to consider how specific programming languages, such as C++, support parallelism through built-in features and libraries. C++ supports multithreaded execution through features such as *std::async*, which play a basic role in parallelism [15]. Other standard multithreading features in C++ that simplify parallel programming include [15]:

- *std::thread*: To manually manage the creation and control of new threads of execution.
- *std::async*: Enables asynchronous function execution, with thread management handled automatically by the system.
- *std::future*: Used with *std::async* or *std::promise* to obtain results from asynchronous executions.
- *std::promise*: For sending a value from one thread to another, typically used with *std::future*.
- *std::mutex*: Prevents race conditions by ensuring that only one thread can access shared data at a time.

Several prior studies have extensively discussed the evolution of multicore architectures, the challenges in optimizing code for parallel execution, and the importance of automatic parallelization to utilize modern hardware efficiently. These works collectively emphasize the necessity of transitioning from sequential to concurrent execution, while highlighting fundamental techniques such as dependency analysis, control flow restructuring, and task distribution across cores. They also explore compiler-assisted approaches, parallel programming models, and language-specific solutions that aim to simplify or automate this transition.

While these contributions have laid important groundwork for parallel computing, particularly in loop and object-oriented contexts, there remains a noticeable gap in the automation of function-level parallelization, especially for statically analyzed, general-purpose C++ applications. This study builds on those foundational insights by introducing a tool that automates the detection, analysis, and transformation of independent function calls into parallel equivalents using C++'s native concurrency libraries. The proposed solution requires no manual intervention or external annotations, offering a fully automated pipeline that strengthens the practical adoption of parallelism in real-world C++ systems.

III. Related Works

Parallel computing has recently become more in demand due to the increasing growth of multi-threaded architecture. There are some papers that discuss certain techniques that aim to efficiently transfer a sequential program into a parallel one.

In [9], the authors introduce an approach that uses the concept of futures synthesis for pure method call parallelization. Their approach employs three strategies to achieve automatic parallelization through future synthesis: parallelism analysis, future synthesis, and threshold synthesis. In the future synthesis phase, the proposed system identifies pure method calls and labels them as async expressions. After that, the system conducts a parallelism analysis to determine which async expressions should be executed sequentially to avoid overhead. Finally, threshold synthesis utilizes the results of the parallelism analysis to synthesize predicate expressions for deciding whether a specific pure method call should be executed in parallel or sequentially.

Our tool focuses on C++ and operates at the static level, targeting functions that are independent and computationally intensive. It does not rely on runtime profiling but instead uses a scoring system to identify which functions should be parallelized. While in study [9] the tool is adaptive and runtime-aware, ours offers a lightweight and deterministic transformation pipeline that produces a ready-to-compile parallel version of the original program, making it ideal for users seeking fast and fully automatic optimization.

In [16], authors introduced Lazy-Parallel Function Calls, a compiler-level approach that enables automatic parallelization in imperative languages by combining lazy evaluation semantics with nano-thread scheduling. Their method ensures correctness by leveraging immutability and deferring function execution until results are required, enabling parallelism without altering program behavior. In contrast, our tool targets eager, function-level parallelization in C++ using *std::async*, automatically identifying independent function calls and transforming them for concurrent execution. While their work emphasizes semantic safety, our approach focuses on performance gains in compiled C++ applications.

In [10], the authors present an automatic technique that analyzes code written in the Java programming language to detect method calls that can be parallelized and transform them accordingly. The proposed technique is based on control dependence and data dependence analysis to construct the execution flow and determine accessed data.

APAC [11] is a compiler that converts sequential source code written in the C++ language into task-based parallelized code by inserting special directives. This compiler utilizes the LibTooling library and clang-tools from the LLVM framework. It encapsulates each function body with an OpenMP directive, and each function or method call with an OpenMP statement.

As authors stated in study [12], there are many recursive functions that contain a significant amount of built-in parallelism, but it is not always clearly visible. Thus, a number of studies have been conducted to detect parallelism and utilize it. Mainly by targeting calls that can be executed concurrently. However, some recursive algorithms do not permit the concurrent execution of function calls. According to the authors [12], they examined ways to extract concurrency from specific recursive function calls that are not typically parallelizable by common parallelization methods. Their primary approach involves parallelizing them at a finer level, relying on multi-threaded architecture and a multi-core as the infrastructure. The authors stated that, compared to conventional parallel systems, multi-core systems support a greater degree of fine-grained parallelism by providing thread parallelism and reducing overhead in thread communication.

Autopar [13] is a tool that targets to automatically parallelize certain recursive function calls using analysis of static programs. It detects the recursive functions inside a certain program. After that, it moves to the analysis phase by gathering related information about such functions without the need for rewriting the program or requiring developer involvement. Finally, Autopar introduces the parallel constructs necessary to achieve automatic parallelization. In fact, their tool applies to a specific class of recursive functions that use a divide-and-conquer approach. They tested four benchmarks—bionic, fractal, heat, and knapsack—using their recursive function parallelization tool.

In [14], the proposed tool applies a coarse-grained approach to task parallelism. It utilizes OpenMP directives and inserts them into the input C code. The output program is a multithreaded C code that can use different cores in a shared memory system. The system operates between the compiler and the application. The generated output is compiled just like any ordinary C code. Both the input and output of their proposed tool are source code, making it a source-to-source transformation tool.

In [17], authors propose two optimizations—Last Parallel Call Optimization (LPCO) and its generalization, Nested Parallel Call Optimization (NPCO) which is designed to improve the efficiency of and-parallel logic programming. Their methods restructure parallel goal execution in logic programs to minimize redundant parallel tasks and enable fast backtracking. These techniques are tailored for the control-parallel execution model, where correctness depends on preserving goal dependencies.

Table I presents a comprehensive comparison of the limitations identified in existing approaches and how our proposed tool addresses these gaps.

TABLE I. LIMITATIONS OF EXISTING APPROACHES AND OUR PROPOSED SOLUTION

Aspect	Existing Approaches Limitations	Our Proposed Tool Solution
Language Support	Limited to specific languages: Java [9][10], C [14], logic programming [17]	Comprehensive C++ support with standard library integration (std::async, std::future)
Function Coverage	Restricted to recursive [12][13] or pure methods [9]; ignores general function calls	Analyzes all function types: recursive, non-recursive, member functions, and standalone functions
Analysis Type	Runtime profiling required [9]; limited static analysis [13][14]	Comprehensive static analysis combining dependency, context, and workload assessment
Automation Level	Semi-automated requiring manual refinement [11]; domain-specific [17]	Fully automated end-to-end pipeline from analysis to code generation

IV. METHODOLOGY

This study introduces a tool that addresses an important parallelization technique: unlike loop parallelization, research on parallelizing function calls is limited. This tool offers an improved method for analyzing and selecting the applicable functions for parallelization in C++ code. This tool includes AST traversal and work potential metrics to evaluate computational effort and identify functions with high parallelization possibility. Moreover, it generates detailed control flow graphs as well as an analysis context in order to give valuable insights to programmers who look to optimize program efficiency through parallel execution. As Fig. 1 shows, the tool has five main functionalities, which we will discuss in detail. These main functionalities are: Function Analysis, Context Analysis for Parallelization, Control Flow Graph (CFG) Generation, Work Potential Calculation, and CFG Visualization.

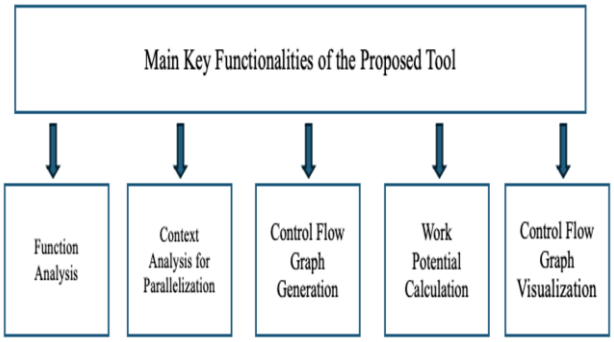


Fig. 1. Five key functionalities of the tool.

Main key functionalities of the proposed tool:

A. Function Analysis

- Parses the C++ source code using Clang's libclang to generate an abstract syntax tree (AST).
- Extracts functions and analyzes each line for function calls to understand dependencies and execution flow.

B. Context Analysis for Parallelization

- Determines whether each function call can be parallelized based on its context (e.g., whether it's inside loops or conditionals).

C. Control Flow Graph (CFG) Generation

- Generates a CFG for each function, including nodes for statements and edges representing control flow.
- Annotates each function call with contextual information (e.g., if part of a loop) to highlight non-parallelizable code.

D. Work Potential Calculation

- Each function is assigned a "work potential" score based on the complexity of its control flow and nested calls.
- Higher work potential indicates a candidate for parallelization, while certain functions (like main) are excluded due to its structural importance.

E. Control Flow Graph (CFG) Visualization

- Uses Graphviz to create a DOT file for each function's CFG, saving it as a DOT, a PDF, and TXT file.
- The CFG graphically represents the control flow, showing how different statements and function calls are interconnected.

Our tool utilizes the use of libclang for parsing C++ code (AST) and graphviz for visualizing CFGs. Libclang is a library that offers tools for parsing and analyzing C++ code. It is part of the LLVM project and enables the generation of an Abstract Syntax Tree (AST). Graphviz, on the other hand, is a visualization software used for creating graphs and diagrams. It is specifically utilized to generate Control Flow Graphs (CFGs), which visually represent the flow of execution within a program, helping programmers understand how different parts of the code are connected and identify potential bottlenecks or parallelization opportunities. Moreover, it usesC++ Parallelization and Synchronization Library (std::async: allows asynchronous function execution and enabling tasks to run concurrently in separate threads). As Fig. 2 illustrates, the tool's architecture starts with C++ source code input, which undergoes scanning to detect functions. This is followed by C++ code analysis, which performs dependency and context analysis generate Control Flow Graphs (CFGs) and calculate work potential. Finally, the tool moves to C++ code generation, completing the process [10].

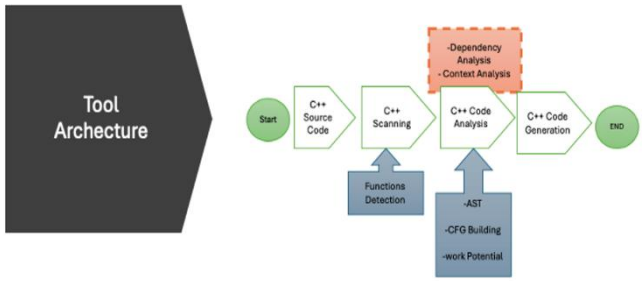


Fig. 2. The proposed tool archeticure.

A detailed description of the tool architecture:

1) Context analysis: This analyzes the context in which function calls occur, meaning it analyzes C++ source code to identify functions that can be parallelized to improve performance.

Tables II and III provide a detailed clarification of the situations under which functions are considered parallelizable and when they are not suitable for parallelization.

2) CFG building: CFG builder drives the analysis of a specified C++ code, generates control flow graphs (CFGs), saves them in a designated output directory, and classifies functions by their work potential (which will be discus1§sed) for advanced parallelization identification.

TABLE II. CONDITIONS FP;OR CONSIDERING A FUNCTION CALL AS POTENTIALLY PARALLELIZABLE

Condition	Reason
Not Part of Control Structure: The function call is not used within the condition of control structures such as if, while, for, or switch statements.	Function calls within conditions are critical for the control flow and may need to be executed sequentially to maintain program correctness
Not Inside Loops: The function call is not located within the body of for or while loops.	Parallelizing function calls inside loops can lead to synchronization issues and may not provide performance benefits due to overhead.
Not Part of Return Statements: The function call is not used directly within a return statement.	The result of the function is immediately required for the return value, so parallelizing it would necessitate waiting for its completion. Attempting to execute such a function asynchronously (in parallel) does not provide any performance improvement and may introduce unnecessary overhead. This essentially cancels out any potential advantages of parallelization in that context.
Not Part of Throw or Assert Statements: The function call is not used within throw or assert statements. **Throw**: is used to signal an exception or error condition. **Assert**: The assert function checks / validates a condition at runtime.	These statements are critical for error handling and program correctness, requiring immediate execution.
Standalone Function Calls: The function call is a standalone statement or assigned to a variable outside the contexts mentioned above.	Such calls can be executed asynchronously without affecting the program's control flow.

TABLE III. CONTEXTS WHERE A FUNCTION CALL IS NOT SUITABLE FOR PARALLELIZATION

Number	Condition
1	Control Structure Conditions
2	Inside Loops
3	Return Statements
4	Part of Throw or Assert Statements

3) Work potential metric: work potential metric is used for prioritizing parallelization by suggesting specific functions to parallelize. This process is based on calculating the computational load or complexity of each function and assigning it a score. Functions with a high work potential score are likely to benefit from parallelization, so the tool identifies and lists these functions at the end of the analysis, see Fig. 3.

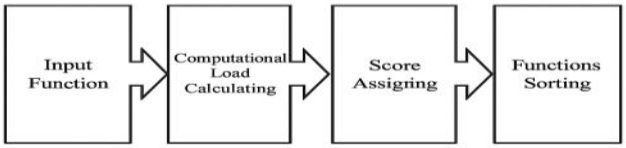


Fig. 3. Steps of the work potential metric.

4) Work potential purposes: The main purposes of the work potential metric are to quantify computational effort because functions with more statements and function calls need more computational processing, and to identify chances for parallelization, as high work potential indicates that a function may benefit from being executed in parallel which enhance performance. However, there are some excluded functions. Certain functions, like main, are excluded from parallelization even if their work potential score is high, because parallelizing the main function does not make sense. It organizes all other functions and controls the entire program's execution, so it is avoided from parallelization.

5) Work potential calculation: Calculation Method for finding Work Potential is a systematic method that requires analyzing C++ functions and utilizing Clang's libclang library to evaluate computational complexity. It uses Clang's libclang library to parse the C++ source code and generate an AST, and traverses each function's AST to analyze its structure. The Work Potential Counter (work_potential) counts statements and function calls. First, it is initialized to zero for each function. Then, it gets incremented for each statement encountered during traversal, as well as for each function call within the function.

V. RESULT AND DISCUSSION

We evaluated the parallelization tool on the BookLibrary benchmark program. BookLibrary, a C++ program, is considered a tool for data processing and analytics of book review data. It includes classes like Review, Book, Author, User, ExtractData, and ExtractDataset in order to organize and analyze information related to books, authors, users, and reviews. Our approach targets programs where function calls account for a significant portion of the workload. Therefore, we chose BookLibrary program because a substantial amount of its sequential work is executed through function calls. BookLibrary.cpp helps users load, analyze, and summarize book review data, with a particular focus on author and user review statistics, enabling insights into authors' popularity and reader engagement. Table IV shows the main components of the C++ code. In addition to those listed in Table IV, there are other functions that calculate and return the most, least, and average-reviewed books, as well as retrieve the most and least-reviewed authors based on their books' reviews.

TABLE IV. OVERVIEW OF MAIN COMPONENTS IN THE C++ CODE

Classification	Description
Data Representation	Review: Represents a review of a book with attributes like book ID and title.
	Book: Represents a book, including its title and list of reviews.
	Author: Represents an author with related books and users.
	User: Represents a user, with a user ID and related reviews.
Data Extraction and Aggregation	ExtractDatase: Reads book and review from CSV files, then stores them .
	ExtractData: create mappings between books and authors, reviews and users, then updates book review data.

A total of 53 functions were automatically extracted from the analyzed source file.

The detected functions range from getBookId, located at line 25, to the main function at the end of the code.

This demonstrates the tool's capability to comprehensively identify and list all functional components for further analysis (see Fig. 4).

```
================================================================================
           Enhanced CFG Builder with Function Call Context Analysis
================================================================================

Analyzing file: books_library.cpp
Found function: getBookId at line 25
Found function: getTitle at line 26
Found function: getUserID at line 27
Found function: getScore at line 29
Found function: getPrice at line 37
Found function: getProfileName at line 38
Found function: getHelpfulness at line 39
Found function: getTime at line 40
Found function: getSummary at line 41
Found function: getText at line 42
Found function: getTitle at line 58
Found function: getAuthors at line 59
Found function: getReviews at line 60
Found function: addReview at line 61
Found function: updateMediumScore at line 62
Found function: getMediumScore at line 76
Found function: resetReviews at line 77
Found function: operator== at line 78
Found function: getDescription at line 81
Found function: getImage at line 82
Found function: getPreviewLink at line 83
Found function: getPublisher at line 84
Found function: getPublishedDate at line 85
Found function: getInfoLink at line 86
Found function: getCategories at line 87
Found function: getRatingCount at line 88
Found function: getFullName at line 100
```

Fig. 4. Extracted function from book_library.cpp.

Fig. 5 below presents the tool's output that demonstrates the automated generation of Control Flow Graphs (CFGs) for several detected functions such as *getBookId*, *getTitle*, *getScore*, *getText*, *getReview*, and *getPrice*. For each function, the tool produces three distinct representations: DOT, PDF, and an extended TXT format to support both visual inspection and structural analysis. This comprehensive output enables the examination of the internal logic and control dependencies within each function. Additionally, it allows tracing execution paths, detecting redundant operations, and better understanding of the relationships among function calls. Overall, this automated CFG generation provides a valuable foundation for program comprehension, debugging, and performance optimization.

```
DOT file for function 'getBookId' saved to cfg_output/getBookId_cfg.dot
PDF file for function 'getBookId' saved to cfg_output/getBookId_cfg.pdf
TXT file for function 'getBookId' saved to cfg_output/getBookId_cfg.txt

Extended CFG for function 'getBookId' saved to cfg_output/getBookId_extended_cfg.pdf
DOT script for extended CFG saved to cfg_output/getBookId_extended_cfg.dot
Extended CFG details for function 'getBookId' saved to cfg_output/getBookId_extended_cfg.txt

DOT file for function 'getTitle' saved to cfg_output/getTitle_cfg.dot
PDF file for function 'getTitle' saved to cfg_output/getTitle_cfg.pdf
TXT file for function 'getTitle' saved to cfg_output/getTitle_cfg.txt

Extended CFG for function 'getTitle' saved to cfg_output/getTitle_extended_cfg.pdf
DOT script for extended CFG saved to cfg_output/getTitle_extended_cfg.dot
Extended CFG details for function 'getTitle' saved to cfg_output/getTitle_extended_cfg.txt

DOT file for function 'getUserID' saved to cfg_output/getUserID_cfg.dot
PDF file for function 'getUserID' saved to cfg_output/getUserID_cfg.pdf
TXT file for function 'getUserID' saved to cfg_output/getUserID_cfg.txt

Extended CFG for function 'getUserID' saved to cfg_output/getUserID_extended_cfg.pdf
DOT script for extended CFG saved to cfg_output/getUserID_extended_cfg.dot
Extended CFG details for function 'getUserID' saved to cfg_output/getUserID_extended_cfg.txt

DOT file for function 'getScore' saved to cfg_output/getScore_cfg.dot
PDF file for function 'getScore' saved to cfg_output/getScore_cfg.pdf
TXT file for function 'getScore' saved to cfg_output/getScore_cfg.txt

Extended CFG for function 'getScore' saved to cfg_output/getScore_extended_cfg.pdf
DOT script for extended CFG saved to cfg_output/getScore_extended_cfg.dot
Extended CFG details for function 'getScore' saved to cfg_output/getScore_extended_cfg.txt

DOT file for function 'getPrice' saved to cfg_output/getPrice_cfg.dot
PDF file for function 'getPrice' saved to cfg_output/getPrice_cfg.pdf
TXT file for function 'getPrice' saved to cfg_output/getPrice_cfg.txt
```

Fig. 5. Control Flow Graph (CFG) files generated for multiple functions.

```
Processing call: {'name': 'resetReviews', 'context': 'Potentially Parallelizable', 'line': 172}
Found call 'resetReviews' at line 172
Processing call: {'name': 'addReview', 'context': 'Potentially Parallelizable', 'line': 177}
Found call 'addReview' at line 177
Processing call: {'name': 'updateMediumScore', 'context': 'Potentially Parallelizable', 'line': 181}
Found call 'updateMediumScore' at line 181
Processing call: {'name': 'updateBooksReviews', 'context': 'Potentially Parallelizable', 'line': 186}
Found call 'updateBooksReviews' at line 186
Processing call: {'name': 'getReviews', 'context': 'Not Parallelizable: Part of return statement', 'line': 189}
Found call 'getReviews' at line 189
```

Fig. 6. Sample tool output showing identified function calls and their parallelization context.

Fig. 6 illustrates the tool's intermediate output during function call analysis. It captures many functions such as resetReviews, addReview, and updateMediumScore, and precise source code line numbers. Each function is reported along with its name, classification context, and source code line number.

```
Parallelizing top 4 functions:
extractMostReviewedAuthor: 11
extractLeastReviewedAuthor: 11
extractAverageReviewedAuthor: 11
getUserForAuthor: 5
CFG generation with function calls completed.
================================================================================
                                Analysis Complete
================================================================================

All CFGs have been generated and saved in the 'cfg_output' directory.
```

Fig. 7. Selected functions for parallelization.

Fig. 7 shows the final summary and highlights the four functions identified as suitable for parallelization based on their work potential score. The functions extractMostReviewedAuthor, extractLeastReviewedAuthor, and extractAverageReviewedAuthor were selected due to their significant contribution to the overall processing load. Control Flow Graph (CFG) generation was completed for all, and the results were saved in the cfg_output directory.

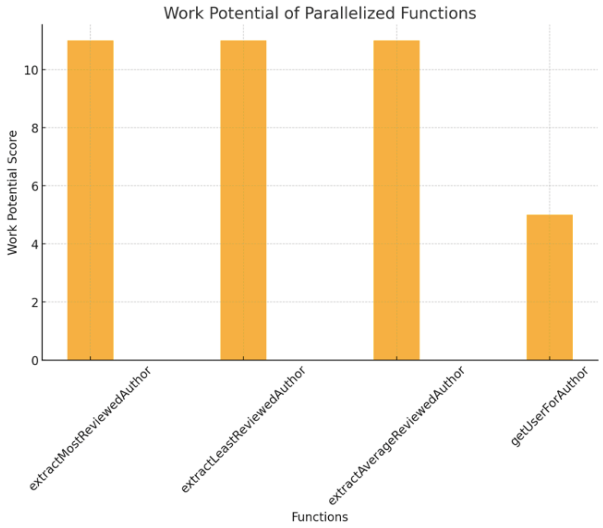


Fig. 8. Work potential scores for selected functions identified for parallelization.

Fig. 8 presents a bar chart showing the work potential scores of selected functions identified by the tool. The x-axis displays four functions: extractMostReviewedAuthor, extractLeastReviewedAuthor,

extractAverageReviewedAuthor, and getUserforAuthor, while the y-axis quantifies their corresponding scores. The first three functions each achieved a high potential score of 11, indicating strong suitability for parallelization due to their computational intensity. In contrast, getUserforAuthor received a lower score of 5, suggesting a relatively smaller parallel workload.

Fig. 9 shows the completion stage of the tool's transformation process, where selected functions have been successfully parallelized. The tool logs the transformed functions: extractMostReviewedAuthor, extractLeastReviewedAuthor, extractAverageReviewedAuthor, and getUserforAuthor and saves the modified source code as books_library_parallelized.cpp.

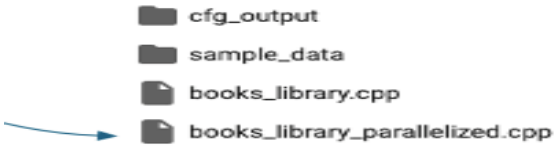


Fig. 9. Generation of the parallelized output file.

Fig. 10. Snapshot of Generated Control Flow Graph (CFG) files.

Fig. 10 presents a visual overview of the directory structure containing the Control Flow Graph (CFG) outputs generated by the tool. For each analyzed function, such as addUser, extractBooks, and extractMostReviewedAuthor, there are many output formats created:

- .dot: the raw DOT graph description format
- .pdf: a visual representation of the control flow
- .txt: a textual report of the CFG
- _extended_ versions: enriched CFGs that include additional semantic or contextual data

These files are organized systematically within the cfg_output directory, enabling developers to inspect the control flow structure of each function in both graphical and textual forms.

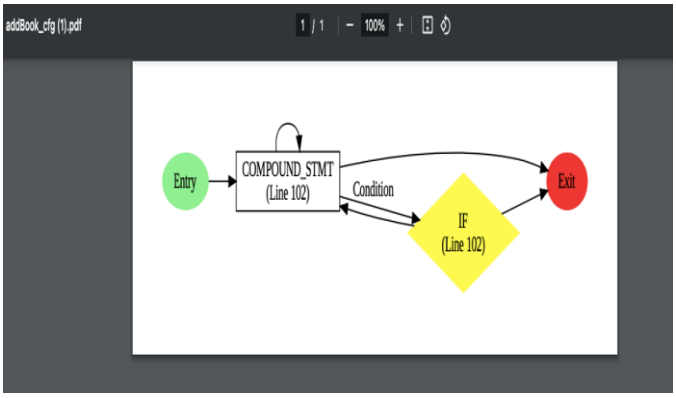


Fig. 11. PDF format Generated Control Flow Graph (CFG) for the Function addBlockFunction.

Fig. 11 presents a sample Control Flow Graph (CFG) generated in PDF format for the selected function addBlock. The graph begins at the Entry node and transitions to a Compound Statement at line 102, indicating a block of grouped instructions. It then leads to a conditional IF node, representing a branching decision in the function's logic.

Fig. 12 represents another format which is DOT format, while Fig. 13 represents TXT format.

Fig. 12. DOT format generated Control Flow Graph (CFG) for the function.

addBook_cfg.dot addBook_cfg.txt

```
Control Flow Graph for Function: addBook
==========================================

Nodes:
  addBook_entry
  addBook_exit
  addBook_stmt_0
  addBook_stmt_0
  addBook_decision_0
  addBook_stmt_0

Edges:
  addBook_entry --> addBook_stmt_0
  addBook_stmt_0 --> addBook_stmt_0
  addBook_stmt_0 --> addBook_decision_0
  addBook_decision_0 --> addBook_stmt_0
  addBook_stmt_0 --> addBook_exit
  addBook_decision_0 --> addBook_exit

```

Fig. 13. TXT format Generated Control Flow Graph (CFG) for the function addBlock( ) function.

```
std::optional<Author> extractMostReviewedAuthor(
    const std::map<std::string, Book>& books,
    const std::multimap<std::string, Review>& reviews) {

    ExtractData extractor(books, reviews);
    std::map<std::string, Author> authors = extractor.getAuthors();
    std::optional<Book> book = extractor.getMostReviewedBook();
```

Fig. 14. Original sequential implementation of extractMostReviewedAuthor function.

Fig. 14 shows the original sequential version of the extractMostReviewedAuthor function, where operations are executed in order. Fig. 15 presents the parallelized version automatically generated by the tool, using std::async to execute getAuthors and getMostReviewedBook concurrently. Our proposed tool identifies these opportunities without user intervention, which enhance the optimization process and facilitate it.

```
std::optional<Author> extractMostReviewedAuthor(
    const std::map<std::string, Book>& books,
    const std::multimap<std::string, Review>& reviews) {

    ExtractData extractor(books, reviews);
    auto authorsTask = std::async(std::launch::async, &ExtractData::getAuthors, &e
    std::map<std::string, Author> authors = authorsTask.get();

    auto bookTask = std::async(std::launch::async, &ExtractData::getMostReviewedBo
    std::optional<Book> book = bookTask.get();
```

Fig. 15. Parallelized Version of extractMostReviewedAuthor Using std::async.

VI. Evaluation Metrics

This section presents a performance analysis and evaluation through comparing serial and parallel implementations, with a focus on improvements achieved via function-level parallelization. The evaluation measures the impact of parallelization across three distinct data processing applications:

A. Book Reviews Analysis System

Processes large datasets of book reviews to extract insights such as the most and least reviewed authors, average review counts, and user-specific data. Ideal for evaluating function-level parallelism in data filtering and aggregation tasks.

B. Employee Performance Analysis System

Analyzes performance metrics of employees to compute scores, rankings, and performance categories. Suitable for testing parallel function execution in systems with structured, repetitive computations.

C. Genomic Sequence Analysis System

Performs operations such as sequence matching, frequency analysis, and data summarization on long character sequences. This benchmark represents workloads with high computational demand and pattern processing.

The result shows the impact of parallel execution in enhancing computational efficiency across different workloads and domains.

In this study, the Improvement metric was selected to quantitatively assess the performance gains achieved through function-level parallelization. The metric calculates the percentage reduction in execution time between the original serial implementation and the parallelized version, using the following formula:

$$\text{Improvement}\ (\%) = [(T_{serial} - T_{parallel}) / T_{serial}] \times 100 \quad (1)$$

Where *Tserial* denotes the execution time of the serial version, and *Tparallel* represents the time taken by the parallel implementation.

D. Book Reviews System Performance Metrics

Table V and Fig. 16 compare the execution times of serial and parallel implementations across various functional components. The results demonstrate that all operations benefited from parallelization, with notable reductions in execution time. The most significant improvement was observed in the Data Extraction function, which executed 43.7% faster compared to the serial version. In contrast, the Least Reviewed Author function showed the smallest improvement at 7.9%. Overall, the total execution time was reduced by 36.1%, confirming that the tool effectively enhanced computational performance.

Fig. 17 presents the performance data in a different format, emphasizing the percentage improvement for each function. Represented as a line graph, this view highlights the relative efficiency gains from parallelization.

TABLE V. BOOK REVIEWS SYSTEM PERFORMANCE METRICS (MILLISECONDS)

FUNCTION	SERIAL EXECUTION TIME	PARALLEL EXECUTION TIME	IMPROVEMENT (%)
DATA EXTRACTION	144.615	81.354	43.7
MOST REVIEWED AUTHOR	21.551	14.312	33.6
LEAST REVIEWED AUTHOR	20.847	19.196	7.9
AVERAGE REVIEWED AUTHOR	24.492	17.826	27.2
USER-AUTHOR ANALYSIS	33.697	23.941	29.0
TOTAL EXECUTION	245.205	156.629	36.1

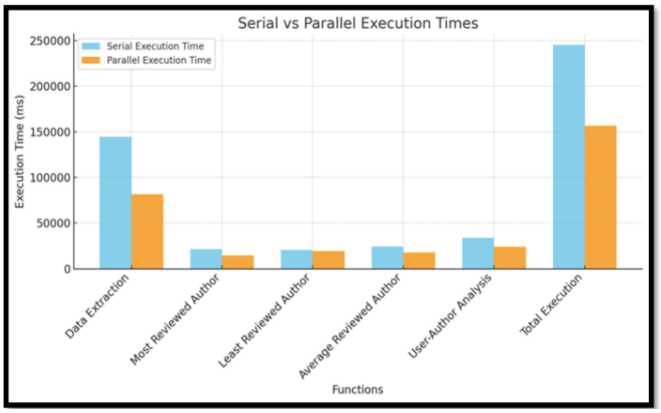


Fig. 16. Serial vs. Parallel execution time (book reviews analysis system).

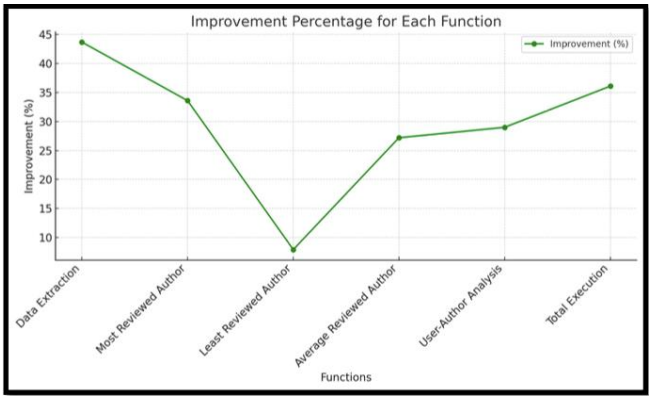


Fig. 17. Improvement percentage across functions (book reviews analysis system).

E. Employee Performance Analysis System

As Table VI, Fig. 18 and Fig. 19 show, most functions show significant performance improvements with parallelization, especially Top Performer Analysis (68.6%) and Data Extraction (64.1%). Skills Analysis also benefited notably (65.8%). However, Department Performance slightly worsened (-3.8%), due to the overhead and the low parallel potential. Overall, total execution time was reduced by 54.6%.

TABLE VI. EMPLOYEES' PERFORMANCE SYSTEM PERFORMANCE METRICS (MILLISECONDS)

Function	Serial Execuion Time	Parallel Execution Time	Improvement (%)
Data Extraction	107.516	38.627	64.1
Top Performer Analysis	34.632	10.873	68.6
Department Performance	17.922	18.611	-3.8
Skills Analysis	485	166	65.8
Department Statistics	15.850	11.766	25.8
Total Execution	176.405	80.049	54.6

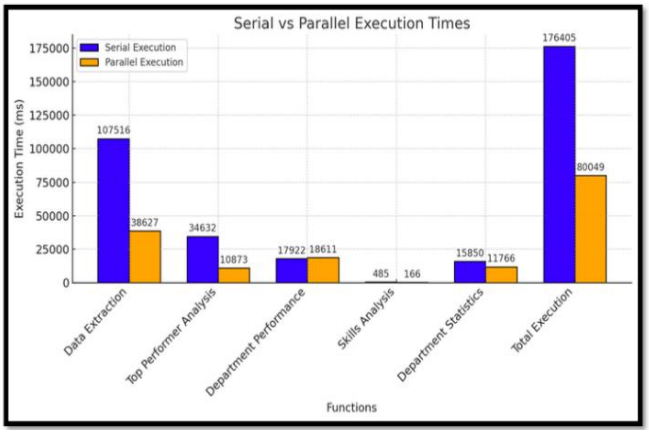


Fig. 18. Serial vs. Parallel execution time (employee performance analysis system).

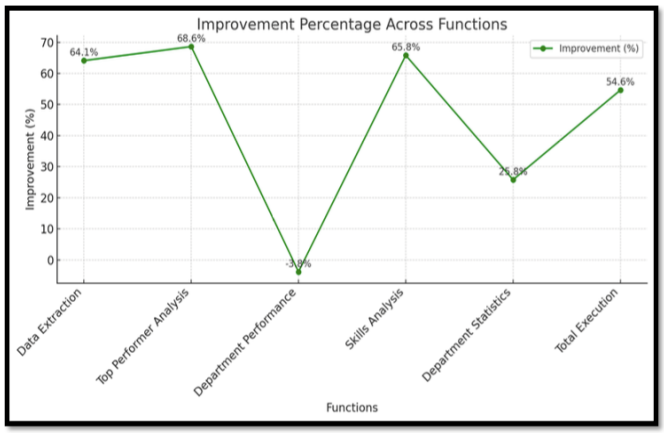


Fig. 19. Improvement percentage across functions (employee performance analysis system).

F. Genomic Sequence Analysis System

As Table VII, Fig. 20 and Fig. 21 show, Data Extraction achieved the highest gain at 54.6%, followed by GC Content Analysis (37%) and Coverage Analysis (30.6%). Overall, total execution time was reduced by 48.3%.

TABLE VII. GENOMIC SEQUENCE SYSTEM PERFORMANCE METRICS (MILLISECONDS)

Function	Serial Execuion Time	Parallel Exection Time	Improvement (%)
Data Extraction	118.149	53.670	54.6
GC Content Analysis	63.252	39.827	37.0
Coverage Analysis	1.902	1.320	30.6
Total Execution	183.303	94.820	48.3

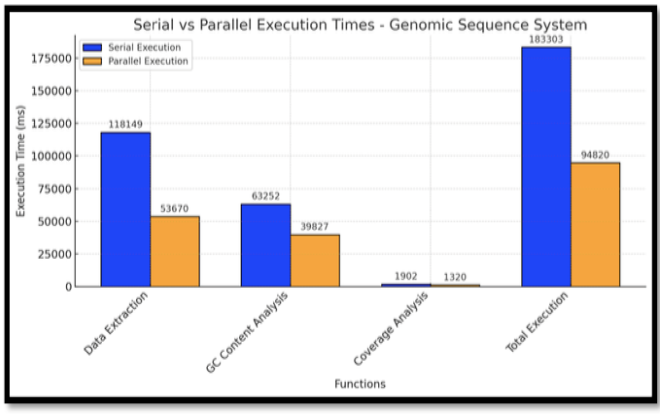


Fig. 20. Serial vs. Parallel execution time (genomic sequence analysis system).

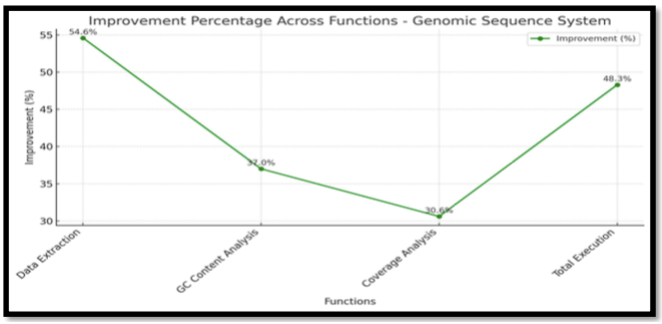


Fig. 21. Improvement percentage across functions (genomic sequence analysis system).

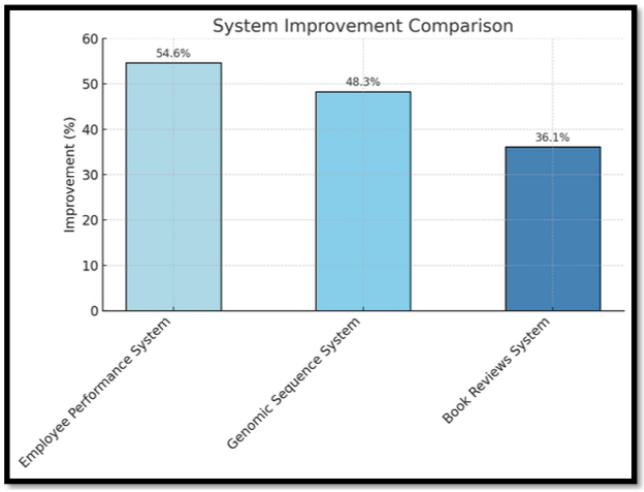


Fig. 22. System improvement comparison.

Fig. 22 presents a bar chart that shows a final comparison of overall system performance improvements achieved through parallelization. The Employee Performance System showed the highest gain at 54.6%, followed by the Genomic Sequence System at 48.3%, while the Book Reviews System achieved a more modest improvement of 36.1%. This confirms the tool's effectiveness across varied application domains. Table VIII provides a comparative overview of our tool and several related tools. The comparison is based on key attributes like parallelization method, analysis type (static or runtime), and language support. [9][12][13].

TABLE VIII. COMPARISON OF OUR TOOL WITH RELATED TOOLS ACROSS VARIOUS ATTRIBUTES

Work	Lazy-Parallel Function Calls	Automatic Parallelization of Pure Method Calls	Fine-Grained Recursive Parallelim	Autopar	OurTool
Languag Focus	Java, C#	Java	C	ANSI C Codes	C++
Optimization Target	Lazy evaluation, closures, nano-threads	Pure method calls	Special Recursive Function Calls	Special Recursive Function Calls	C++ source code, Function calls based on three fundament-al analyses
Automation Level	Partial; compiler-assisted	automatic, compiler-based	automatic	automatic	Fully automated pipeline: detects, classifies, and transforms suitable function calls to parallel equivalents.

VII. CONCLUSION

This paper presents a tool that successfully automates function-level parallelization in C++ through a pipeline that includes dependency analysis, context analysis, and workload assessment. These analyses enable the identification of independent, high-workload function calls that can be safely parallelized without modifying the original program logic. The tool also generates detailed Control Flow Graphs (CFGs) for each function in DOT, PDF, and TXT formats, offering valuable insights into code structure and parallelization opportunities.

Experimental evaluations on three benchmark programs demonstrated consistent improvements in execution efficiency. Performance gains were particularly notable in compute-intensive functions such as data extraction and analysis. The Employee Performance System achieved the highest

improvement at 54.6%, followed by the Genomic Sequence System at 48.3%, and the Book Reviews System at 36.1%. These results confirm the effectiveness of the tool's analysis and transformation strategies and underscore its contribution to advancing practical, automated parallelization in performance-critical C++ applications.

Future work may focus on extending the tool's capabilities to support parallelization of additional code constructs and the use of MPI, enabling distributed-memory execution and scalability across multiple nodes. Such an extension would broaden the tool's applicability beyond shared-memory systems.

REFERENCES

[1] R. C. O. Rocha, L. F. W. Góes, and F. M. Q. Pereira, "Automatic parallelization of recursive functions with rewriting rules," *Sci. Comput. Program.*, vol. 173, pp. 128–152, 2019.

[2] C. Navarro, N. Hitschfeld, and L. Mateu, "A survey on parallel computing and its applications in data-parallel problems using GPU architectures," *Commun. Comput. Phys.*, vol. 15, pp. 285–329, 2013.

[3] J. Parkhurst, J. Darringer, and B. Grundmann, "From single core to multi-core: Preparing for a new exponential," in *Proc. Int. Conf. Comput.-Aided Design (ICCAD)*, 2006, pp. 67–72.

[4] D. Geer, "Industry trends: Chip makers turn to multicore processors," *Computer*, vol. 38, no. 5, pp. 11–13, 2005.

[5] K. Sujatha, P. V. N. Rao, A. A. Rao, V. G. Sastry, V. Praneeta, and R. K. Bharat, "Multicore parallel processing concepts for effective sorting and searching," in *Proc. PACES-2015, Dept. of ECE, KL Univ.*, 2015.

[6] M. Sohal and R. Kaur, "Automatic parallelization: A review," *Int. J. Comput. Sci. Mobile Comput. (IJCSMC)*, vol. 5, no. 5, pp. 17–21, 2016.

[7] J. Kwiatkowski and R. Iwaszyn, "Automatic program parallelization for multicore processors," in *Lecture Notes in Computer Science*, vol. 6081, pp. 236–245, 2010.

[8] A. Barve, S. Khomane, B. Kulkarni, S. Ghadage, and S. Katare, "Parallelism in C++ programs targeting objects," in *Proc. Int. Conf. Adv. Comput., Commun. Control (ICAC3)*, 2017, pp. 1–6.

[9] R. Surendran and V. Sarkar, "Automatic parallelization of pure method calls via conditional future synthesis," Rice Univ., Houston, TX, USA, 2011.

[10] A. Midolo and E. Tramontana, "An automatic transformer from sequential to parallel Java code," Univ. of Catania, Italy, 2023.

[11] G. Kusoglu, B. Bramas, and S. Genaud, "Automatic task-based parallelization of C++ applications by source-to-source transformations," in *Compas 2020: Parallélisme/Architecture/Système/Temps Réel*, Lyon, France, 2020.

[12] D. Saougkos, A. Mastoras, and G. Manis, "Fine-grained parallelism in recursive function calls," Univ. of Ioannina, Greece, 2012.

[13] M. E. Kalender, C. Mergenci, and O. Ozturk, "AutopaR: An automatic parallelization tool for recursive calls," Dept. Comput. Eng., Bilkent Univ., Turkey, 2014.

[14] M. Mathews and J. P. Abraham, "Implementing coarse-grained task parallelism using OpenMP," Dept. Comput. Sci. Eng., Mar Athanasius Coll. of Eng., India, 2015.

[15] P. Mehta, S. Singh, D. Roy, and M. M. Sarma, "Comparative study of multi-threading libraries to fully utilize multi-processor/multi-core systems," *Int. J. Current Eng. Technol.*, 2014.

[16] S. S. Chatterjee and R. Gururaj, "Lazy-parallel function calls for automatic parallelization," in *Proc. Int. Conf. Comput. Intell. Inf. Technol. (CIIT)*, vol. CCIS 250, pp. 811–816, Springer, 2011.

[17] E. Pontelli and G. Gupta, "Last parallel call optimization and its generalization," *J. Logic Program.*, vol. 27, no. 1, pp. 1–43, 1996.

Unveiling the Drivers of Consumer Purchase Intention in Short-Form Video Marketing

Merisa Syafrina, Viany Utami Tjhin
Information System Management Department, Bina Nusantara University, Jakarta, Indonesia 11480

***Abstract*—Short-video features on e-commerce platforms have become a key driver of social commerce, enhancing user engagement and purchase intention. However, user reviews of Shopee Video reveal issues such as disruptive autoplay, limited content control, and unintuitive navigation. While prior studies have examined engagement and satisfaction in general e-commerce, limited research has explored how short-video features within social commerce influence purchase intention through user engagement. This study fills that gap by analysing the factors affecting consumer behaviour toward Shopee Video. Sentiment analysis of user reviews identified common dissatisfaction themes, followed by a quantitative survey of 300 Shopee Video users in Indonesia. Using Structural Equation Modeling with the Partial Least Squares (SEM-PLS) approach, the results show that user engagement significantly mediates the relationship between system quality, information quality, and technology adoption and usefulness on purchase intention. The study extends existing models to the social commerce context, providing insights for optimizing short-video features to strengthen engagement and conversion.**

Keywords*—*E-commerce; purchase intention; Shopee Video; social commerce; SEM-PLS

I. Introduction

Digital technology occupies an increasingly important position in today's world and has undeniably become a strong social trend closely linked to the need for building a sustainable future [1]. Technological advances have also brought new changes, where most people now rely on internet-based digital technology in their daily lives [2]. According to the Indonesian Internet Service Providers Association (APJII), in 2024 around 221 million out of 278 million Indonesians were already connected to the internet, with 89.44% of the population accessing it through smartphones. The use of the internet for business has also continued to rise since the 1980s, reaching its peak in the 1990s with the emergence of online marketplace platforms, commonly known as electronic commerce (e-commerce), such as eBay and Amazon [3]. Since then, business activities involving interactions between companies and clients have become faster through the internet, which enables transactions to be conducted online. Moreover, access to e-commerce platforms has become increasingly easier anytime and anywhere through mobile devices [4]. E-commerce has also made a positive impact on Indonesia's economy by increasing entrepreneurs' income while providing convenience for consumers to shop without leaving their homes [5].

In Indonesia, several e-commerce platforms have grown rapidly, such as Shopee, Tokopedia, Bukalapak, Lazada, and Blibli. Among them, Shopee is the most popular platform [6]. Launched in Singapore in 2015 as a Consumer-to-Consumer (C2C) marketplace, Shopee has since evolved into a hybrid model combining C2C and Business-to-Consumer (B2C) through Shopee Mall, which hosts well-known brand stores. According to SimilarWeb data in early 2024, Shopee ranked first among e-commerce competitors in Indonesia with 235.9 million visits in February 2024, followed by Tokopedia (100.3 million), Lazada (45.03 million), and Blibli (24.15 million). The consistently high user interest has driven Shopee to continuously innovate with new features and strategies to maintain its competitive advantage, including discount programs, Shopee coins, diverse payment options, campaigns, livestreaming, and other interactive features [7][8]. In line with this innovation, Shopee introduced Shopee Video, a short-video feature that allows users to better understand products before purchasing. Based on a 2024 Ipsos study on online shopping experience and satisfaction, Shopee was rated as the e-commerce platform with the best short-video experience, outperforming TikTok Shop, Tokopedia, and Lazada. The survey involved 506 respondents across Indonesia aged 17–45, with an equal proportion of men and women. Shopee ranked first with a 61% satisfaction level in short-video experiences, followed by TikTok Shop (18%), Tokopedia (10%), and Lazada (9%) [9].

The short-video format on platforms like Shopee typically ranges from 10 to 30 seconds, allowing users to quickly understand product information, making the experience both efficient and engaging [10]. Such videos often cover diverse topics, from fashion trends and lifestyle entertainment to commercial advertisements and e-commerce promotions [11]. By adopting features like Shopee Video, Shopee has embraced the social commerce model, which integrates social interaction with online shopping [12]. This model allows users to actively engage through liking products, writing reviews, sharing experiences, and directly purchasing within the same platform [13]. Consequently, short-video formats with embedded purchase links and livestreaming functions have become integral to modern consumption behavior, enabling consumers to "socialize while shopping" more conveniently.

However, despite its popularity, user feedback reveals persistent challenges. Preliminary scraping of 33 Apple App Store reviews and 288 Google Play Store reviews identified complaints such as disruptive autoplay, lack of content control, and unintuitive navigation [14]. These usability and content-related issues highlight that, while short-video features may enhance engagement, they may not always lead to positive purchase intentions.

Previous studies on e-commerce and short-video use have primarily focused on user satisfaction, entertainment value, and technology adoption in general contexts [15][16]. Yet, few have examined how short-video features within social commerce ecosystems influence purchase intention through user engagement mechanisms. Moreover, existing research tends to overlook how system quality, information quality, and technology adoption interact with engagement behavior to shape consumer purchase decisions. This represents a critical research gap, as the short-video environment involves richer interaction, user-generated content, and algorithm-driven experiences that differ from conventional e-commerce.

Therefore, this study aims to fill that gap by extending the Delone and McLean Information System Success Model with constructs relevant to short-video and social commerce contexts, specifically integrating User Engagement as a mediating variable between System Quality, Information Quality, Technology Adoption and Usefulness, and Purchase Intention. This theoretical contribution provides empirical validation of established models in a new cultural and technological setting, Indonesia's social commerce landscape and offers practical insights for optimizing user-centered video-based features.

A. Research Questions

Based on the above background, this study formulates the following research questions:

- What factors influence consumer purchase intention in using the short-video feature (Shopee Video) on the Shopee e-commerce platform?
- How does each of these factors affect consumer purchase intention toward using the short-video feature (Shopee Video) on the Shopee e-commerce platform?

B. Research Objectives

The primary objective of this study is to identify and analyze the factors that influence consumer purchase intention in using the short-video feature, Shopee Video, within the Shopee e-commerce application. Specifically, the research aims to examine how system quality, information quality, and technology adoption and usefulness shape consumer behavior through the mediating role of user engagement. By understanding these relationships, the study seeks to provide empirical insights into how short-video features within social commerce platforms can effectively enhance user engagement and drive purchasing decisions. Furthermore, this research intends to contribute both theoretically and practically by enriching the literature on social commerce and offering recommendations for e-commerce practitioners and digital marketers to optimize short-video features that align with user expectations and behavioral patterns.

II. Literature Review

A. Previous Studies on Short-Video Features in E-Commerce and Social Commerce

Short videos have become a dominant form of digital content, shaping consumer behavior and purchase decisions across e-commerce and social commerce platforms. In e-commerce, short videos combine product information with engaging visual storytelling, facilitating consumers' transition from browsing to purchasing. Studies emphasize that product demonstration and visual aesthetics in short videos significantly influence viewing duration, suggesting that engaging content enhances user attention. However, this study primarily focused on content attributes, not on user satisfaction or behavioural outcomes such as purchase intention [17].

Similarly, explored persona perception and shared values in social commerce short videos, showing that perceived user identity positively affects purchase intention. Yet, their model emphasized social factors and emotional alignment, overlooking system and information quality aspects that may affect user experience. Other works, examined dual pathways of rationality and affection in livestream-based social commerce but focused more on social atmosphere than on short-video usability or engagement mechanisms [14].

In contrast, recent research analysed impulse buying and consumer satisfaction on platforms like Shopee Video and TikTok [18][19]. Their findings reinforced the role of emotional arousal and pleasure but were limited to Generation Z samples and focused on impulsive or affective responses rather than sustained engagement or cognitive evaluation. Hence, while prior studies have demonstrated the persuasive potential of short videos, most focus on content and emotional stimuli, leaving a gap in understanding how system, information, and technological aspects interact with user engagement to drive purchase intention especially within e-commerce-based short videos like Shopee Video rather than purely social platforms.

B. System Quality

System quality, as conceptualized in the DeLone and McLean Information Systems Success Model, measures the reliability, stability, and responsiveness of an information system. Prior studies showed that platform performance and seamless navigation influence user stickiness and loyalty in video-based platforms like Douyin [20]. However, such works focused on overall experience rather than feature-specific performance. In the Shopee Video context, system quality directly affects how users interact with video features, such as playback stability, autoplay control, and navigation. This study extends the D&M model by validating system quality's effect within a video-based social commerce feature, emphasizing its indirect influence through user engagement.

C. Information Quality

Information quality involves the relevance, clarity, and trustworthiness of product-related information conveyed through short videos. It was found that vivid and diagnostic content enhances trust and user commitment [20], whereas other research identified that well-demonstrated product information sustains user attention [17]. Despite these insights, few studies have analyzed how information quality in short-video formats shapes purchase intention through user engagement, particularly when users rely on short, rapidly consumed content for decision-making. This study contributes by examining how users perceived information quality from

Shopee Video influences engagement and subsequent purchasing behavior.

D. Technology Adoption and Usefulness

Technology adoption and perceived usefulness reflect how users evaluate the benefits and ease of using a system. Past research widely applied the Technology Acceptance Model (TAM) to e-commerce contexts, linking perceived usefulness to satisfaction and purchase intention. However, few have tested it within the social commerce short-video environment, where users engage through both consumption and interaction. By integrating TAM into the D&M model, this study explores how perceived usefulness affects engagement behavior and purchase intention in short-video shopping experiences, expanding TAM's application into a more interactive and entertainment-driven context [16].

E. User Engagement

User engagement represents users' cognitive, emotional, and behavioral involvement with digital platforms. It has been shown to mediate the relationship between technology-related factors and purchase behaviors [19]. Yet, many prior studies treated engagement as an outcome variable rather than a mediating mechanism. This study positions user engagement as a key mediator that links system quality, information quality, and technology adoption to purchase intention. By doing so, it extends the D&M model into the social commerce domain, offering a novel perspective on how interactive engagement behaviors transform functional experiences into purchase motivation.

F. Purchase Intention

Purchase intention reflects consumers' willingness to buy after interacting with a product or feature. Prior works confirmed that emotional and social cues affect purchase intention in social commerce, but few have tested its determinants in video-based e-commerce platforms like Shopee [14], [21]. This study addresses that gap by combining both technical (system and information quality) and behavioral (engagement, usefulness) factors, presenting an integrated view of how short-video experiences influence purchase decisions.

III. Methodology

A. Research Design

This study employs a quantitative research design to examine the factors influencing purchase intention through user engagement in the context of Shopee Video, a short-video feature within a social commerce platform. The research process began with a preliminary study using scraping methods to identify user complaints and behavioral patterns from app store reviews. Insights from this stage were used to formulate the research problem and develop a conceptual model grounded in relevant theories and previous studies. Data were collected using an online survey distributed via social media to Shopee Video users in Indonesia aged 17 years and above see Table V and Table VI of Appendix A.

To visualize the research process, the overall study workflow is illustrated in Fig. 1, titled Overall Study Process. This diagram outlines the sequential steps: (1) preliminary research through sentiment analysis of app reviews, (2) literature review and model development, (3) hypothesis formulation, (4) data collection, (5) data analysis using SEM-PLS and (6) conclusion.

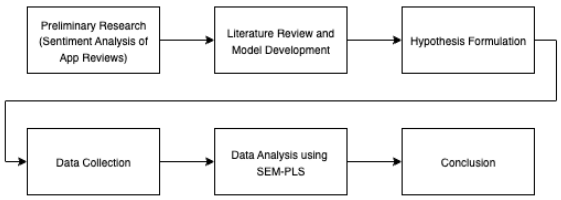


Fig. 1. Overall study process.

B. Sampling

The population of this study comprises Shopee users in Indonesia who have used the Shopee Video feature. Because the exact number of active users of this feature is unknown, a simple random sampling technique was applied. Data were collected using a structured online questionnaire distributed through Google Forms. Following the recommendation of previous study for SEM-PLS, the minimum sample size was determined to be 300 respondents calculated as 100 times the number of independent variables [22] (system quality, information quality, and technology adoption & usefulness) to ensure adequate statistical power and model reliability.

$$n\ minimum = \text{Independent variables} \times 100$$

$$= 3 \times 100$$

$$= 300 \text{ respondents}$$

C. Research Model

The conceptual model (Fig. 2) integrates constructs from the Information Systems Success Model (DeLone & McLean, 2016), the Technology Acceptance Model (TAM), and literature on social commerce engagement. The model posits that System Quality (SQ), Information Quality (IQ), and Technology Adoption & Usefulness (TAU) influence Purchase Intention (PI) both directly and indirectly through User Engagement (UE) as a mediating variable.

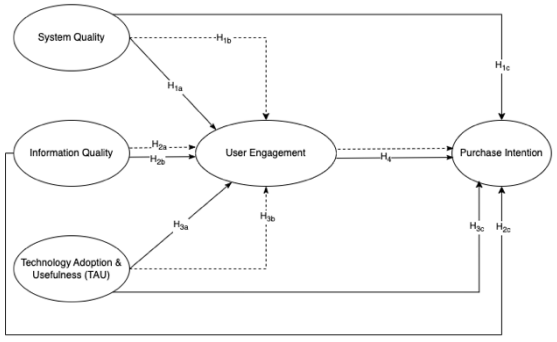


Fig. 2. Conceptual model.

This framework extends prior studies in social commerce by emphasizing the mediating role of engagement a dimension often underexplored in existing models. While prior research examined satisfaction or perceived usefulness, this study highlights engagement as a psychological bridge connecting quality perceptions and behavioral outcomes. Thus, the proposed model not only replicates but also advances previous

approaches by integrating user-centered insights from short-video platform behavior.

D. Research Hypotheses

Based on the conceptual model, the following hypotheses were formulated:

H1a: System Quality significantly influences User Engagement.

H1b: System Quality significantly influences Purchase Intention through User Engagement.

H1c: System Quality significantly influences Purchase Intention.

H2a: Information Quality significantly influences User Engagement.

H2b: Information Quality significantly influences Purchase Intention through User Engagement.

H2c: Information Quality significantly influences Purchase Intention.

H3a: Technology Adoption & Usefulness significantly influences User Engagement.

H3b: Technology Adoption & Usefulness significantly influences Purchase Intention through User Engagement.

H3c: Technology Adoption & Usefulness significantly influences Purchase Intention.

H4a: User Engagement significantly influences Purchase Intention.

E. Data Analysis

Data were analyzed using the Structural Equation Modeling-Partial Least Squares (SEM-PLS) approach with SmartPLS software. This method was chosen for its suitability in testing complex models with multiple constructs and mediating relationships. The analysis process included validity and reliability testing, evaluation of R-square values, and hypothesis testing to confirm the proposed relationships among variables.

F. Study Limitations

This study is limited to the short-video feature, Shopee Video, within the Shopee e-commerce application. The research focuses on users in Indonesia who have experience using Shopee Video, specifically those aged 17 years and above and residing in the Jabodetabek area. The study adopts and modifies research models from previous studies related to system quality, information quality, technology adoption, user engagement, and purchase intention. As such, the findings may not be fully generalizable to users in other regions or to different e-commerce platforms. Future research could expand the sample coverage and explore comparative analyses across multiple social commerce platforms to strengthen external validity.

IV. Data Analysis

A total of 310 questionnaires were collected. After data verification, 10 responses were excluded for not meeting the screening requirement of being Shopee Video users. Therefore, 300 valid responses were used for further analysis. This yields a usable response rate of 96.77 percent, confirming that the sample is adequate for the subsequent statistical analysis using the SEM-PLS method.

A. Gender Distribution

Among the 300 respondents, 222 were female (74 percent) and 78 were male (26 percent). This finding indicates that Shopee Video's user base is predominantly female, consistent with broader trends in Indonesia's e-commerce and social commerce platforms, where women are generally more active in online shopping and product content interaction.

B. Age Group Distribution

Respondents were categorized into four generational groups based on Indonesia's Internet Service Provider Association (APJII) classification. Generation Z (aged 17–27 years) constituted the largest group with 195 respondents (65 percent), followed by Millennials aged 28–43 years (69 respondents or 23 percent), Generation X aged 44–58 years (26 respondents or 8.6 percent), and Baby Boomers aged 59 years and above (10 respondents or 3.33 percent). The dominance of Generation Z and Millennials highlights the strong engagement of younger users with short-video content and their increasing role in shaping social commerce trends in Indonesia.

C. Occupation

The largest occupational group consisted of private employees (121 respondents or 40.33 percent), followed by students (111 respondents or 37 percent), freelancers (25 respondents or 8.33 percent), homemakers (18 respondents or 6 percent), and entrepreneurs (13 respondents or 4.33 percent). A small proportion were government employees (8 respondents or 2.66 percent) and retirees (4 respondents or 1.33 percent). These results show that the Shopee Video audience includes a wide range of working adults and students, groups most familiar with digital shopping and short-form video consumption.

D. Duration of Shopee App Usage

Most respondents had used the Shopee application for more than three years (222 respondents or 74 percent). Sixty respondents (20 percent) had used it for one to three years, while smaller groups reported six months to one year (11 respondents or 3.66 percent) and less than six months (7 respondents or 2.33 percent). This long-term usage pattern indicates that respondents have sufficient experience to assess the quality and usefulness of the Shopee Video feature.

E. Primary Purpose of Using Shopee Video

When asked about their main purpose for using Shopee Video, 188 respondents (62.66 percent) indicated that they primarily watch product reviews or user testimonials. Seventy-one respondents (23.66 percent) use it to find special promotions, 28 respondents (9.33 percent) seek shopping inspiration, and 13 respondents (4.33 percent) use it mainly for entertainment. This suggests that Shopee Video is primarily perceived as an informational and promotional tool rather than an entertainment platform.

F. Product Categories Frequently Purchased via Shopee Video

Fashion-related items, including apparel, footwear, and accessories, were the most frequently purchased products (176 respondents or 58.66 percent). The next categories were Beauty and Personal Care (62 respondents or 20.66 percent), Electronics and Gadgets (26 respondents or 8.66 percent), Home Goods and Furnishings (21 respondents or 7 percent), Food and Beverages (11 respondents or 3.66 percent), and Books and Magazines (4 respondents or 1.33 percent). The dominance of fashion products aligns with prior research showing that visual and aesthetic content in short videos strongly influences consumer purchase intention in lifestyle-related product categories.

G. Familiarity with Shopee Video

Most respondents reported being familiar (170 respondents or 56.66 percent) or very familiar (110 respondents or 36.66 percent) with the Shopee Video feature, while only 20 respondents (6.66 percent) indicated limited familiarity. The high familiarity rate suggests that users have substantial exposure to the feature and can provide reliable feedback regarding system quality, content relevance, and engagement experience.

H. Frequency of Shopee Video Usage Per Week

The highest frequency category was one to two times per week (135 respondents or 45 percent), followed by less than once a week (67 respondents or 22.33 percent), three to five times per week (67 respondents or 22.33 percent), and more than five times per week (31 respondents or 10.33 percent). These results indicate moderate engagement levels among users, suggesting that Shopee Video functions as a complementary browsing and product discovery tool rather than a daily-use feature.

I. Average Viewing Duration Per Session

Most respondents (136 respondents or 45.33 percent) reported spending less than five minutes per session watching Shopee Videos. Another 129 respondents (43 percent) viewed content for five to fifteen minutes, while smaller groups watched for sixteen to thirty minutes (24 respondents or 8 percent) or more than thirty minutes (11 respondents or 3.66 percent). This short viewing duration aligns with the core purpose of short-video features in e-commerce: delivering concise, product-oriented content that facilitates rapid purchase decisions.

J. Purchase Behavior via Shopee Video

A large majority of respondents (265 respondents or 88.33 percent) had made at least one purchase after watching Shopee Video, whereas 35 respondents (11.66 percent) had not. This finding underscores the feature's effectiveness in converting viewers into buyers, reinforcing its strategic importance in Shopee's social commerce ecosystem. The high purchase conversion rate also validates the relevance of this study's research model, which investigates how system quality, content quality, technology adoption, and user engagement jointly influence purchase intention.

K. Outer Model Analysis

1) Validity test: Validity was assessed using the Average Variance Extracted (AVE) to confirm the degree to which each construct explains its corresponding indicators. An AVE value above 0.5 indicates that the construct captures more than half of the variance of its observed variables. The initial analysis showed all constructs exceeded the 0.5 threshold, although three indicators SQ2 (System Stability), IQ5 (Freedom from Tribe, Religion, Race, and Inter-group relations-related elements), and PI4 (Seeking Additional Product Information) had loading values below 0.7. These items were removed to improve the model's validity. After the refinement, the AVE values for all constructs improved and met the validity requirement as seen in Table I.

TABLE I. DISCRIMINANT VALIDITY

Variable	Average Variance Extracted (AVE)	Conclusion
System Quality	0.562	Valid
Information Quality	0.698	Valid
Technology Adoption and Usefulness	0.715	Valid
User Engagement	0.654	Valid
Purchase Intention	0.762	Valid

Following indicator elimination, the AVE values were System Quality (0.562), Information Quality (0.698), Technology Adoption and Usefulness (0.715), User Engagement (0.654), and Purchase Intention (0.762). As all values exceeded 0.5, the constructs demonstrated convergent validity.

2) Reliability test: Reliability measures the internal consistency of items within each construct. The Composite Reliability (CR) and Cronbach's Alpha coefficients were used for this purpose, with values above 0.7 considered acceptable. The results in Table II indicate that all constructs met or exceeded the reliability threshold. System Quality (CR=0.774; α=0.748), Information Quality (CR=0.857; α=0.855), Technology Adoption and Usefulness (CR=0.810; α=0.802), User Engagement (CR=0.901; α=0.894), and Purchase Intention (CR=0.850; α=0.842). This confirms that each construct consistently measures its intended concept. Fig. 3 shows Cronbach's alpha representation.

TABLE II. RELIABILITY TEST RESULT

Variable	Composite Reliability	Cronbach's Alpha	Conclusion
System Quality	0.774	0.748	Reliable
Information Quality	0.857	0.855	Reliable
Technology Adoption and Usefulness	0.810	0.802	Reliable
User Engagement	0.901	0.894	Reliable
Purchase Intention	0.850	0.842	Reliable

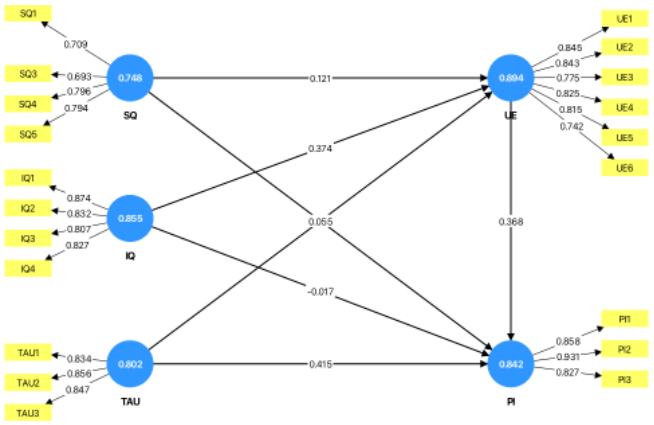


Fig. 3. Cronbach's alpha representation.

In summary, the measurement model fulfils all validity and reliability requirements (Fig. 4), supporting its use in further structural model testing.

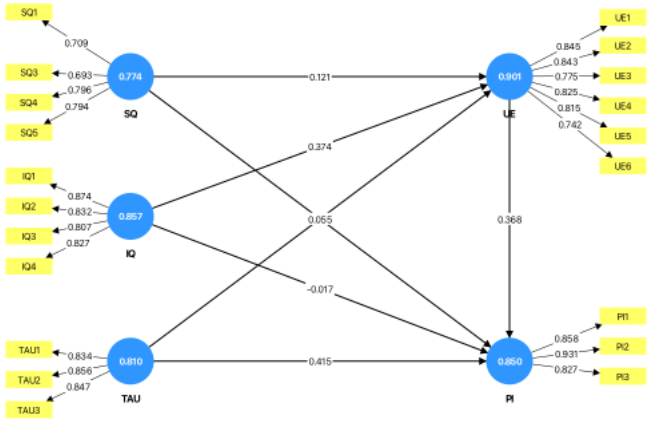


Fig. 4. Composite reliability representation.

L. Inner Model Analysis

1) The coefficient of determination or R-square (R²): The R-square (R²) test measures how well independent variables explain the variance in dependent variables. R² values of 0.67, 0.33, and 0.19 indicate substantial, moderate, and weak levels of explanatory power, respectively. As shown in the results, the R² value for User Engagement was 0.514, indicating that 51.4 percent of its variance is explained by System Quality, Information Quality, and Technology Adoption and Usefulness. Meanwhile, the R² value for Purchase Intention was 0.409, meaning that 40.9 percent of the variance in purchase intention is explained by the independent variables and the mediating role of user engagement. As shown in Table III, R² values fall within the moderate category, suggesting that the model adequately captures the factors influencing user engagement and purchase intention toward Shopee Video.

TABLE III. R-SQUARE RESULTS FOR LATENT VARIABLES

Variable	R²	R² Adjusted	Conclusion
User Engagement	0.514	0.507	Moderate
Purchase Intention	0.409	0.403	Moderate

M. Hypothesis Test

The hypothesis testing was conducted using the bootstrapping method in SmartPLS with a significance level of 0.05. The analysis shown in Table IV examined both direct and indirect effects among constructs based on the proposed hypotheses (H1a–H4a).

TABLE IV. HYPOTHESIS TESTING RESULT

Correlation	O	T	P	Conclusion
SQ → UE	0.121	1.550	0.121	Not Significant
SQ → UE → PI	0.045	1.552	0.128	Not Significant
SQ → PI	0.055	0.965	0.335	Not Significant
IQ → UE	0.374	5.326	0.000	Significant
IQ → UE → PI	0.138	4.028	0.000	Significant
IQ → PI	-0.017	0.239	0.811	Not Significant
TAU → UE	0.218	2.624	0.009	Significant
TAU → UE → PI	0.080	2.335	0.020	Significant
TAU → PI	0.415	6.129	0.000	Significant
UE → PI	0.368	6.446	0.000	Significant

H1a: System Quality → User Engagement. The path from System Quality to User Engagement yielded a t-value of 1.550 (p=0.121), which is below the critical value of 1.97. Therefore, H1a is rejected, indicating that system quality does not significantly influence user engagement. This suggests that although users may expect stable and responsive performance, engagement is more strongly driven by other factors such as information relevance and perceived usefulness.

H1b: System Quality → Purchase Intention (Mediated by User Engagement). The indirect effect of System Quality on Purchase Intention through User Engagement produced a t-value of 1.552 (p=0.128), leading to the rejection of H1b. This result implies that system stability alone does not meaningfully enhance purchase intention, even when user engagement is considered as a mediator.

H1c: System Quality → Purchase Intention. The direct relationship between System Quality and Purchase Intention was not significant (t=0.965, p=0.335). Thus, H1c is rejected. These findings collectively indicate that while a technically sound system is necessary, it is insufficient on its own to drive purchasing behavior in short-video-based e-commerce environments.

H2a: Information Quality → User Engagement. Information Quality had a significant positive impact on User Engagement, with a t-value of 5.326 (p=0.000). Therefore, H2a is accepted. High-quality, accurate, and relevant information increases user trust and stimulates interaction, making it a key driver of engagement within Shopee Video.

H2b: Information Quality → Purchase Intention (Mediated by User Engagement). The indirect path from Information Quality to Purchase Intention via User Engagement was also significant (t=4.028, p=0.000), supporting H2b. This indicates that users who perceive the video content as informative and

trustworthy are more likely to form positive engagement experiences that enhance their purchase intentions.

H2c: Information Quality → Purchase Intention. The direct relationship between Information Quality and Purchase Intention was not significant (t=0.239, p=0.811). Thus, H2c is rejected. Although users value accurate information, it primarily affects purchase intention indirectly through engagement rather than as a standalone factor.

H3a: Technology Adoption and Usefulness → User Engagement. Technology Adoption and Usefulness had a significant positive effect on User Engagement (t=2.624, p=0.009). Therefore, H3a is accepted. This shows that users who perceive Shopee Video as easy to use and beneficial are more likely to engage with it actively.

H3b: Technology Adoption and Usefulness → Purchase Intention (Mediated by User Engagement). The mediating effect of User Engagement on the relationship between Technology Adoption and Usefulness and Purchase Intention was significant (t=2.335, p=0.020), leading to the acceptance of H3b. This indicates that user engagement acts as a crucial psychological mechanism that translates perceived usefulness into actual purchase intention.

H3c: Technology Adoption and Usefulness → Purchase Intention. A strong and significant direct relationship was found between Technology Adoption and Usefulness and Purchase Intention (t=6.129, p=0.000). Thus, H3c is accepted. This result reinforces the importance of technological ease and functional benefits in encouraging consumers to make purchases after watching short videos.

H4a: User Engagement → Purchase Intention. User Engagement significantly influenced Purchase Intention (t=6.446, p=0.000), confirming H4a. This highlights that engagement—through emotional connection, interaction, and trust—plays a pivotal role in shaping consumers' buying decisions within the Shopee Video ecosystem.

Overall, six hypotheses were supported while four were rejected. The significant pathways emphasize the mediating role of user engagement and the strong influence of technology adoption and information quality on consumers' purchasing decisions.

V. Result and Discussion

A. Result

The structural model analysis using SEM-PLS confirmed that not all hypothesized relationships among the five variables were statistically significant. Among the tested variables, Information Quality, Technology Adoption and Usefulness, and User Engagement demonstrated strong positive relationships, while System Quality showed no significant effect on engagement or purchase intention. Specifically, Information Quality had a significant positive influence on User Engagement ($\beta = 0.374$, $p < 0.001$), while Technology Adoption and Usefulness positively affected both User Engagement ($\beta = 0.218$, $p = 0.009$) and Purchase Intention ($\beta = 0.415$, $p < 0.001$). Furthermore, User Engagement itself exhibited a strong positive influence on Purchase Intention ($\beta = 0.368$, $p < 0.001$), confirming its mediating role within the model.

Conversely, System Quality showed no significant direct or indirect effect on either User Engagement or Purchase Intention ($p > 0.05$). This suggests that users' perceptions of system performance—such as loading speed or interface stability—do not substantially drive engagement or purchasing behavior in this context. Instead, informational value and perceived usefulness of technology play a more central role in motivating user behavior within Shopee Video.

The coefficient of determination (R^2) analysis revealed that 51.4% of the variance in User Engagement and 40.9% of the variance in Purchase Intention could be explained by the model. These values indicate a moderate explanatory power, suggesting that while the included variables capture key behavioral influences, other contextual or psychological factors may also contribute to purchase intentions in social commerce environments.

B. Discussion

The findings reinforce the growing importance of social commerce mechanisms, where content-driven engagement mediates the relationship between technology and consumer behavior in shaping online purchase decisions.

The insignificant relationship between System Quality and both User Engagement and Purchase Intention contrasts with earlier studies which emphasized that system performance and stability enhance user experience [23] [24]. This discrepancy may be explained by contextual differences. In the Indonesian market, users of Shopee Video often access the platform via mobile networks with varying speeds, making system performance a baseline expectation rather than a differentiating factor. Hence, engagement depends less on system reliability and more on the perceived usefulness and emotional appeal of content.

In contrast, Information Quality significantly influenced User Engagement, who found that clear, vivid, and diagnostic product information builds consumer trust and stimulates emotional connection [21]. In the Shopee Video context, users engage more with videos that demonstrate authenticity and provide practical product insights rather than polished promotional material.

Technology Adoption and Usefulness emerged as a strong driver of both engagement and purchase intention. Where perceived usefulness directly enhances behavioral intention in video-based commerce [25]. This finding highlights that users engage with short-video features when they perceive them as convenient and beneficial tools for informed decision-making.

The mediating role of User Engagement confirms its theoretical significance as a psychological bridge between functional attributes and behavioral outcomes. Engagement amplifies the effects of both Information Quality and perceived usefulness, validating prior assertions that engagement is not merely a behavioral response but a motivational mechanism in social commerce [21].

Collectively, these findings reveal that purchase intention in short-video–based e-commerce is driven more by content

relevance and perceived value than by technical system attributes. This insight contributes to extending the Information Systems Success Model (DeLone & McLean) into a social commerce context by introducing user engagement as a central mediating construct that links technology-related and information-related factors to behavioral intention.

C. Practical Implications

1) Enhancing information quality: Shopee should prioritize creating authentic, concise, and informative product videos to build user trust and engagement. Seller education and AI-based content moderation could improve diagnosticity and reduce misleading claims.

2) Optimizing engagement features: Interactive elements such as real-time comments, likes, or loyalty rewards can deepen emotional and behavioral engagement, which directly influences purchase decisions.

3) Strengthening technology adoption: Beyond improving system performance, Shopee can emphasize usability, convenience, and personalization in its short-video interface to reinforce the perceived usefulness of the feature.

4) Cultural context adaptation: Platform strategies should reflect Indonesian users' preference for social validation and relatable content, as engagement is shaped by social interaction rather than pure functionality.

VI. Conclusion and Recommendation

A. Conclusion

This study investigated the key factors influencing consumer purchase intention in using short-video features within e-commerce platforms, focusing on Shopee Video as a case study. The results demonstrate that Information Quality, Technology Adoption and Usefulness, and User Engagement significantly influence consumer purchase intention, while System Quality does not show a significant direct or mediated effect.

User Engagement plays a crucial mediating role, linking informational and technological attributes to purchase behavior. This highlights engagement as the primary psychological pathway through which users translate perceived value into actual purchase intentions. Overall, the study extends the Delone and McLean Information Systems Success Model by validating its applicability in a social commerce context, where interactive and emotionally engaging features redefine how consumers experience online shopping.

Beyond Shopee, this research establishes that short-video features in e-commerce platforms shape consumer behavior through engagement and perceived usefulness rather than purely through technical performance. It underscores that the success of video-based commerce depends on the balance between informational authenticity, emotional resonance, and social interactivity.

B. Recommendation

1) Cross-platform comparison: Future research should examine similar short-video commerce features across different e-commerce platforms (e.g., TikTok Shop, Lazada) to compare user engagement patterns and validate model generalizability across markets.

2) Extended variables: Future models could incorporate constructs such as Perceived Ease of Use, Perceived Enjoyment, or Information Security to capture broader behavioral and psychological dimensions influencing purchase intention.

3) Longitudinal analysis: Conducting longitudinal studies would help track evolving user engagement and purchase behavior as short-video features mature, offering insights into retention and loyalty over time.

4) Integration of social and emotional dimensions: Further exploration into social influence, trust, and community identity could deepen understanding of emotional factors that drive engagement in video-based commerce.

5) Practical innovation: Shopee and similar platforms can leverage these insights to refine engagement algorithms, improve recommendation relevance, and create personalized video experiences that foster stronger purchase intentions and long-term user loyalty.

Acknowledgment

Data supporting this study are openly available from Mendeley Data at https://doi.org/10.17632/6kdsysrc89.1, which includes the respondents' survey data. All analytical results are fully presented in the tables and figures within this paper.

Merisa Syafrina contributed to the conceptualization, methodology, data collection, analysis, and writing of the manuscript. Viany Utami Tjhin contributed to supervision, project coordination, and manuscript review. Both authors have read and approved the final version of this article and are accountable for all aspects of the work.

References

[1] O. Lingaitienė, G. Byčenkaitė, and A. Burinskienė, "Application of digital technologies for sustainable business development," Sep. 2024. doi: 10.3846/bm.2024.1199.

[2] W. A. Harsanto, N. Matondang, and R. P. Wibowo, "The Use of Technology Acceptance Model (TAM) to Analyze Consumer Acceptance Towards E-Commerce Websites. A Case of the Plantage.id Digital Transformation Solution," J. Environ. Dev. Stud., vol. 4, no. 2, pp. 206–213, Sep. 2023, doi: 10.32734/jeds.v4i2.13144.

[3] B. Abdul-Raheem and M. K. Ajape, "Integrating the Sunnah Practice into the Operation of E-Commerce," AL-BURHĀN J. QURʾĀN SUNNAH Stud., vol. 8, no. 2, Aug. 2024, doi: 10.31436/alburhn.v8i2.348.

[4] B. Taneja, "The Digital Edge for M-Commerce to Replace E-Commerce," 2021, pp. 299–318. doi: 10.4018/978-1-7998-8587-0.ch016.

[5] C. W. Prasetyandari, "E-Commerce as Indonesia's Economic Development Effort," Indones. J. Econ. Manag., vol. 3, no. 1, pp. 70–78, Nov. 2022, doi: 10.35313/ijem.v3i1.4456.

[6] E. F. Komariyah, "The Impact of Online Satisfaction and Trust in Building Loyalty," Int. J. Business, Law, Educ., vol. 5, no. 2, pp. 2158–2164, Aug. 2024, doi: 10.56442/ijble.v5i2.824.

[7] E. M. Asih, "Analisis pada Shopee sebagai E-Commerce Terpopuler di Indonesia," J. Ekon. Bisnis Antart., vol. 2, no. 1, pp. 73–79, Jun. 2024, doi: 10.70052/jeba.v2i1.299.

[8] N. Safirah and N. Anggrainie, "Impact of Digital Marketing, Influencer, Live Features, and Reviews Shopee Purchase With Brand Image," Int. J.

Econ., vol. 3, no. 2, pp. 1124–1130, Oct. 2024, doi: 10.55299/ijec.v3i2.1065.

[9] A. Yesidora, "Riset: Video Pendek Produk di Shopee Unggul Ketimbang TikTok Shop." Accessed: Nov. 11, 2024. [Online]. Available: https://katadata.co.id/digital/e-commerce/672314c7665d0/riset-video-pendek-produk-di-shopee-unggul-ketimbang-tiktok-shop

[10] M. Rozaq and R. U. Nugrahani, "Penggunaan Platform Video Pendek Sebagai Strategi Komunikasi Pemasaran Digital untuk UMKM," J. Komun. Nusant., vol. 5, no. 1, pp. 21–30, Jun. 2023, doi: 10.33366/jkn.v5i1.271.

[11] J. Feng, "Review of Research on E-commerce Short Videos," Front. Humanit. Soc. Sci., vol. 4, no. 4, pp. 235–242, Apr. 2024, doi: 10.54691/at122k44.

[12] J.-R. Fu, I.-W. Lu, J. H. F. Chen, and C.-K. Farn, "Investigating consumers' online social shopping intention: An information processing perspective," Int. J. Inf. Manage., vol. 54, p. 102189, Oct. 2020, doi: 10.1016/j.ijinfomgt.2020.102189.

[13] H. Han, H. Xu, and H. Chen, "Social commerce: A systematic review and data synthesis," Electron. Commer. Res. Appl., vol. 30, pp. 38–50, Jul. 2018, doi: 10.1016/j.elerap.2018.05.005.

[14] M. Deng, Y. Yang, and B. Sun, "Research on the Purchase Intention of Social Commerce Consumers in Video Streams: Dual Pathways of Affection and Rationality," Behav. Sci. (Basel)., vol. 14, no. 9, p. 738, Aug. 2024, doi: 10.3390/bs14090738.

[15] N. K. S. Suastiari and L. P. Mahyuni, "Determinants of E-Commerce User Satisfaction: the Mediating Role of Purchase Intention," J. Ekon. Bisnis dan Kewirausahaan, vol. 11, no. 3, p. 345, 2022, doi: 10.26418/jebik.v11i3.56329.

[16] D. K. Pramudito, A. Titin Sumarni, E. Diah Astuti, B. Aditi, and Magdalena, "The Influence of User Trust and Experience On User Satisfaction Of E-Commerce Applications During Transactions in Mini Markets Using Delon and McLean Method," J. Sistim Inf. dan Teknol., vol. 5, no. 4, pp. 1–7, Oct. 2023, doi: 10.60083/jsisfotek.v5i4.310.

[17] Y. Guo et al., "Analyzing and Predicting Consumer Response to Short Videos in E-Commerce," ACM Trans. Manag. Inf. Syst., 2024, doi: 10.1145/3690393.

[18] T. T. A. Ngo, H. L. T. Nguyen, H. P. Nguyen, H. T. A. Mai, T. H. T. Mai, and P. L. Hoang, "A comprehensive study on factors influencing online impulse buying behavior: Evidence from Shopee video platform," Heliyon, vol. 10, no. 15, p. e35743, 2024, doi: 10.1016/j.heliyon.2024.e35743.

[19] N. K. Hoi and L. K. Yin, "Short Videos, Big Decisions: A Preliminary Study of Tik Tok's Role in E-Commerce Consumer Behaviour," Eur. J. Bus. Manag. Res., vol. 8, no. 3, pp. 72–81, 2023, doi: 10.24018/ejbmr.2023.8.3.1951.

[20] Q. Yang and Y. C. Lee, "What Drives the Digital Customer Experience and Customer Loyalty in Mobile Short-Form Video Shopping? Evidence from Douyin (TikTok)," Sustain., vol. 14, no. 17, 2022, doi: 10.3390/su141710890.

[21] X. Shen and J. Wang, "How short video marketing influences purchase intention in social commerce: the role of users' persona perception, shared values, and individual-level factors," Humanit. Soc. Sci. Commun., vol. 11, no. 1, pp. 1–13, 2024, doi: 10.1057/s41599-024-02808-w.

[22] J. F. H. Jr., L. M. Matthews, R. L. Matthews, and M. Sarstedt, "PLS-SEM or CB-SEM: updated guidelines on which method to use," Int. J. Multivar. Data Anal., vol. 1, no. 2, p. 107, 2017, doi: 10.1504/IJMDA.2017.087624.

[23] W. H. DeLone and E. R. McLean, Information Systems Success Measurement, vol. 2, no. 1. 2016. doi: 10.1561/2900000005.

[24] M. Al Kofahi, H. Hassan, and R. Mohamad, "DeLone and McLean information systems success model: a literature review," Int. J. Bus. Inf. Syst., vol. 48, no. 4, pp. 452–481, 2025, doi: 10.1504/IJBIS.2025.145551.

[25] C. He and J. Yang, "The Influence of Mobile Short Video Content Marketing on Consumers' Purchase Intention: A Test of Two-Mediation Model," Acad. J. Bus. Manag., vol. 5, no. 19, 2023, doi: 10.25236/AJBM.2023.051916.

[26] M. Deng, Y. Yang, and B. Sun, "Research on the Purchase Intention of Social Commerce Consumers in Video Streams: Dual Pathways of Affection and Rationality," Behav. Sci. (Basel)., vol. 14, no. 9, p. 738, 2024, doi: 10.3390/bs14090738.

APPENDIX A

Table V presents the constructs used in the research model, along with their corresponding measurement codes, indicators, and reference sources. This table serves as a structured overview that links each latent variable with its specific measurement items, providing clarity and transparency regarding the operationalization of the constructs.

TABLE V. CONSTRUCT AND MEASUREMENT ITEMS

Variable	Indicators	Code	References
System Quality	Short video loading speed	SQ1	[21]
	System stability	SQ2	[26]
	Minimal disruptions (glitch/lag)	SQ3	
	Ease of navigation within the short video feature	SQ4	[19]
	System responsiveness to user interactions	SQ5	[26]
Information Quality	Clarity of product information in short videos	IQ1	[19]
	Consistency of messages conveyed by video creators	IQ2	[21]
	Trust in product claims made in short videos	IQ3	[21]
	Completeness of the product information presented	IQ4	[26]
Technology Adoption & Usefulness	Ease of using the short video feature	TAU1	[19]
	Usefulness of product demonstrations in short videos	TAU2	[19]
	Trust in the seller's credibility influencing purchase decisions	TAU3	[21]
User Engagement	Emotional connection with the seller	UE1	[26]
	Sense of direct interaction through the video	UE2	[26]

	Level of trust built through engagement in short videos	UE3	[21]
	Frequency of interactions in short videos (like)	UE4	[21]
	Frequency of interactions in short videos (share)	UE5	
	Frequency of interactions in short videos (comment)	UE6	
Purchase Intention	Intention to purchase after watching short videos	PI1	[19]
	Willingness to consider purchasing after watching short videos	PI2	[21]
	Likelihood of seeking more product information after watching short videos	PI3	[19]

Table VI provides the complete set of questionnaire items used in the survey. Each construct item was measured using a 7-point Likert scale (1 = Strongly Disagree, 2 = Disagree, 3 = Somewhat Disagree, 4 = Neutral, 5 = Somewhat Agree, 6 = Agree, 7 = Strongly Agree) to capture respondents' level of agreement with each statement. In addition, demographic and experience-related questions were included to support descriptive analysis but were not incorporated into the structural model.

TABLE VI. SURVEY QUESTIONS

Section	Survey Questions	Response Options / Scale
Profile Questions	Gender	Male, Female
	Age	17-27 years, 28-43 years, 44-58 years, ≥59 years
	Occupation	Students, Private Employee, Government Employee, Freelancer, Entrepreneur, Homemaker, Informal Sector Worker (e.g. online driver, daily laborer, etc.), Retirees
	Duration of Shopee App Usage	<6 months, 6 months – 1 year, 1-3 years, >3 years
Information about User Experience of Shopee Video	Primary Purpose of Using Shopee Video	Product reviews or user testimonials, shopping inspiration, find special promotions, entertainment
	Product Categories Frequently Purchased via Shopee Video	Fashion, beauty and personal care, electronics and gadget, home and furnishings, food and beverages, books and magazines
	Familiarity with Shopee Video	Very familiar, familiar, less familiar
	Frequency of Shopee Video Usage per Week	<1 time, 1-2 times, 3-5 times, >5 times
	Average Viewing Duration per Session	<5 minutes, 5-15 minutes, 16-30 minutes, >30 minutes
	Purchase Behavior via Shopee Video	Yes, No
Construct Measurement Items	Shopee Video loads quickly without having to wait long.	7-point Liket scale
	Shopee Video rarely experiences technical issues when I use it. Issues or errors include the app not opening, videos not playing smoothly, or features not responding.	
	Shopee Video runs smoothly without interruptions such as stuttering or lag.	
	I can easily find and watch the videos I want on Shopee.	
	When I like, comment on, or share a video link, the Shopee Video system responds quickly.	
	The content on Shopee Video clearly explains the product details.	
	The information presented in Shopee Video is consistent from beginning to end.	
	I believe that the product claims in Shopee Video are accurate and truthful.	
	Shopee Video provides complete information about the products being sold.	
	Shopee Video is free from elements of ethnicity, religion, race, and inter-group issues, violence, or inappropriate content.	
	I find the Shopee Video feature easy to use and understand.	
	Shopee Video helps me understand how a product works or how to use it.	
	I tend to trust sellers on Shopee Video when they recommend products.	
	I feel emotionally connected to the sellers on Shopee Video.	
	When watching Shopee Video, I feel as if I am directly interacting with the seller.	

	I trust the products more after engaging in interactions on Shopee Video (for example, through comments or likes).	
	I often interact on Shopee Video by giving likes.	
	I often interact on Shopee Video by leaving comments.	
	I often interact on Shopee Video by sharing video links with others.	
	I immediately purchase a product after seeing the content on Shopee Video.	
	I feel motivated to buy a product after watching Shopee Video.	

Integrative Hybrid Metaheuristic Algorithm for Hyperparameter Optimisation in Pre-Trained Convolutional Neural Network Models (I-HAHO)

Nazleeni Samiha Haron, Jafreezal Jaafar, Izzatdin Abdul Aziz, Mohd Hilmi Hasan, Muhammad Hamza Azam
Department of Computing, Universiti Teknologi Petronas, Perak, Malaysia

***Abstract*—Hyperparameter optimisation (HPO) remains a fundamental challenge in deep learning, especially for pre-trained convolutional neural networks (CNNs). While pre-trained models reduce the computational burden of training from scratch, their effectiveness depends heavily on tuning parameters such as learning rate, batch size, dropout, weight decay, and optimizer type. The search space of hyperparameters is large, nonlinear, and highly dataset-dependent, making traditional techniques like grid search, random search, and Bayesian optimisation insufficient. This paper introduces I-HAHO, an Integrative Hybrid Metaheuristic Algorithm that combines Artificial Bee Colony (ABC) for global exploration and Harris Hawks Optimisation (HHO) for local exploitation. A diversity-based phase-switching mechanism dynamically regulates exploration and exploitation, allowing the optimiser to adapt its search behaviour to varying landscape conditions. Experiments on CIFAR-10, CIFAR-100, SVHN, and TinyImageNet with three CNN architectures (VGG16, ResNet50, EfficientNet-B0) demonstrate up to 6.9% accuracy improvements. I-HAHO enhances adaptability, scalability, and robustness for hyperparameter tuning.**

Keywords*—Hyperparameter Optimisation (HPO); Convolutional Neural Networks (CNNs); Artificial Bee Colony (ABC); Harris Hawks Optimisation (HHO); Hybrid Metaheuristic Algorithm

I. Introduction

Convolutional Neural Networks (CNNs) have become indispensable in modern computer vision, enabling breakthroughs in tasks such as classification, object detection, and medical imaging. Pre-trained models like VGG16, ResNet50, and EfficientNet-B0, trained on large datasets such as ImageNet, provide transferable feature representations that accelerate training and improve performance, especially when target datasets are small [1], [2].

Despite their success, the performance of pre-trained CNNs is highly sensitive to hyperparameter settings such as learning rate, batch size, dropout rate, and optimizer type. Selecting these parameters is challenging because the search space is high-dimensional, non-convex, and dataset-dependent, making manual tuning and traditional search strategies inefficient. Techniques such as grid search and random search require extensive computational effort, while Bayesian optimisation often struggles with scalability and mixed-variable types [3], [4].

To address these challenges, metaheuristic algorithms have been widely adopted for hyperparameter optimisation [5], [6]. Their ability to balance exploration (global search for promising regions) and exploitation (local refinement near good solutions) makes them well-suited for navigating complex search spaces [7],[8]. Among these, the Artificial Bee Colony (ABC) and Harris Hawks Optimisation (HHO) algorithms have shown promise in deep learning optimisation tasks [9], [10]. However, both algorithms have limitations: ABC often lacks strong exploitation, while HHO may converge prematurely in multi-modal landscapes [11], [12].

This motivates the development of hybrid approaches that combine complementary strengths. In this study, we propose I-HAHO (Integrative Hybrid ABC-HHO), which unifies ABC's exploration capability with HHO's exploitative refinement. A dynamic phase-switching mechanism guided by population diversity regulates the balance between exploration and exploitation, allowing the optimiser to adapt its search behaviour to varying landscape conditions, preventing premature convergence and preserving diversity [13], [14].

The main contributions of this paper can be summarized as follows. First, we propose a novel hybrid metaheuristic algorithm, I-HAHO, which integrates the global exploration strength of ABC with the local exploitation capability of HHO to achieve robust hyperparameter optimisation for pre-trained CNNs. Second, we introduce a diversity-driven dynamic phase-switching strategy that adaptively balances exploration and exploitation during the optimisation process, preventing premature convergence and preserving solution diversity. Third, we provide an extensive evaluation of I-HAHO across multiple benchmark datasets, including CIFAR-10, CIFAR-100, SVHN, and Tiny ImageNet, applied to three widely used CNN architectures: VGG16, ResNet50, and EfficientNet-B0. Finally, we demonstrate that I-HAHO achieves substantial accuracy improvements of up to 6.9% compared with grid search, random search, Bayesian optimisation, and standalone metaheuristics, highlighting its effectiveness, scalability, and generalizability for deep learning optimisation tasks.

The remainder of this manuscript is organized as follows. Section II presents the theoretical background of pre-trained CNN architectures and the hyperparameter optimisation problem, followed by a review of recent metaheuristic and hybrid approaches in the Section III. Section IV introduces the proposed I-HAHO algorithm, detailing its design, modifications to ABC and HHO, and the diversity-driven phase-switching mechanism. Section V describes the experimental setup, including datasets, CNN models, hyperparameter search spaces, and preprocessing strategies. It also presents and analyses the

results, comparing I-HAHO with conventional and metaheuristic baselines in terms of accuracy and convergence. It discusses the implications, strengths, limitations, and potential extensions of the proposed approach. Finally, Section VI concludes the paper by highlighting key findings and suggesting directions for future research.

II. Background of Study

This section provides a discussion of relevant theoretical foundations and an introduction to the hyperparameter optimization problem.

A. Pre-Trained Models

Deep learning has revolutionized computer vision, with Convolutional Neural Networks (CNNs) emerging as the foundation of state-of-the-art models across a wide range of applications, including medical imaging, object recognition, and autonomous driving [15]. CNNs learn hierarchical feature representations directly from raw data, eliminating the need for handcrafted feature extraction and enabling scalable solutions for high-dimensional inputs such as images.

However, training CNNs from scratch is computationally expensive and requires large labelled datasets. To address these limitations, pre-trained CNNs such as VGG16, ResNet50, and EfficientNet-B0 are commonly adopted in transfer learning. These models, initially trained on large datasets like ImageNet, serve as general-purpose feature extractors and can be fine-tuned for domain-specific tasks.

1) VGG16: Introduced by [16], is characterized by its simplicity, uniform 3×3 convolution filters, and deep sequential layers. While powerful in feature extraction, its large parameter size (~138M) makes it computationally demanding. Fig. 1 depicts the VGG16 architecture.

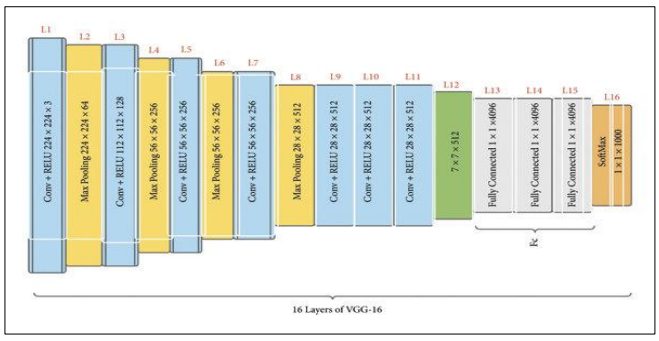


Fig. 1. VGG16 architecture [16].

2) ResNet50: Developed by [17], incorporates residual connections that enable the effective training of deeper architectures by mitigating vanishing gradient problems. This design has significantly improved CNN scalability and accuracy in challenging tasks [18]. Fig. 2 depicts the ResNet50 architecture.

3) EfficientNet-BO: Proposed by [20], introduces a compound scaling strategy that jointly scales depth, width, and resolution. It achieves high accuracy with a lightweight parameter count (~5.3M), making it suitable for mobile and embedded applications [21] (Şafak & Barışçı, 2024). Fig. 3 depicts the EfficientNet-B0 architecture.

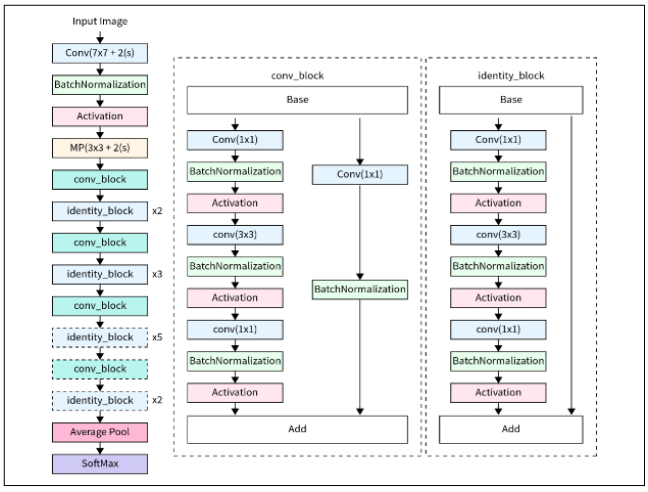


Fig. 2. ResNet50 architecture [19].

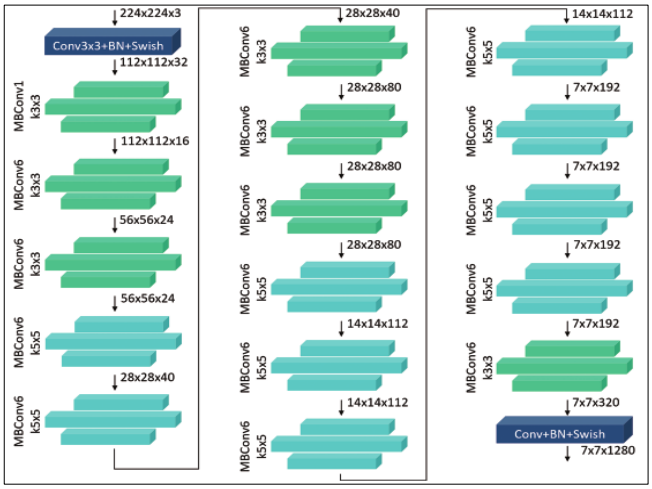


Fig. 3. EfficientNet-B0 architecture [20].

4) Comparison among pre-trained CNN models or architectures: To validate the proposed HPO strategies and algorithms in real-world scenarios, three pre-trained CNN architectures (VGG16, ResNet50, and EfficientNet-B0) are selected for experimental evaluation. These models represent a range of design philosophies and computational complexities, allowing for comprehensive testing against the three core performance objectives: accuracy, adaptability, and computational efficiency. Table I summarises the characteristics of the chosen pre-trained CNN models.

B. Hyperparameter Optimisation (HPO)

The primary goal of HPO is to automate the hyperparameter tuning process to ensure optimal performance of ML models on specific tasks. The motivation for employing HPO lies in its ability to address challenges posed by manual tuning, such as non-linear hyperparameter interactions and the computational burden of evaluating large search spaces.

TABLE I. CHARACTERISTICS OF CHOSEN PRE-TRAINED CNN MODELS

Criteria	VGG16	ResNet50	EfficientNet-B0
Architecture Depth	16 layers (13 convolutional + 3 fully connected)	50 layers with residual connections	82 layers (scalable depth, width, and resolution)
Key Innovation	Sequential architecture with uniform 3×3 convolutions	Residual connections to solve vanishing gradient problem	Compound scaling method to balance network dimensions
Model Size	~138 million parameters	~25.5 million parameters	~5.3 million parameters
Computational Complexity	High due to depth and fully connected layers (~15.4 GFLOPs)	Moderate (~3.9 GFLOPs)	Low (~0.39 GFLOPs)
Real-World Applications	Medical imaging, video classification, and object detection	Autonomous vehicles, semantic segmentation, and face recognition	Mobile vision, IoT devices, and scalable vision applications

In the context of convolutional neural networks (CNNs), hyperparameters play a pivotal role in shaping the architecture, learning dynamics, and generalization ability. These settings, determined prior to training, influence feature extraction, overfitting prevention, and convergence behaviour. CNN hyperparameters can generally be categorized into three main groups [22], [23].

- Hyperparameters of convolutional layers (i.e., number of convolutional layers, number of kernels, size of kernels, stride, padding, dilation rate, activation functions, and filter initialization methods).
- Hyperparameters of fully connected layers (i.e., dropout rate, connectivity pattern, number of neurons, activation functions, weight regularization, and initialization methods).
- General Hyperparameters (i.e., batch size, learning rate, learning rate decay, optimizer type, number of epochs, momentum, and gradient clipping).

1) HPO problem formulation: HPO is the process of identifying the optimal hyperparameters for a machine learning model to maximize its performance on a given task. This is formulated as an optimisation problem where the objective is to find the hyperparameter configuration that minimizes a loss function or maximizes an evaluation [24], [25].

a) General formulation

Let,

$\lambda = \lambda_1, \lambda_2, \ldots, \lambda_n$: The hyperparameter vector, where each λ_i represents a specific hyperparameter.

$M(\lambda)$: The machine learning model trained with hyperparameter configuration ƎvƛƐ.

$L(M(\lambda), D)$: The loss function of the model on dataset $\|1\mathcal{D} = \{(\mathcal{D}\}_{train}, D_{val})$ where D train and D val are the training and validation datasets, respectively.

The goal of HPO is to find the optimal hyperparameter configuration ƎvƛƐ such that:

$$\lambda^* = argmin_{\lambda \in \Lambda} L(M(\lambda), D_{val})$$

where Λ is the search space of hyperparameters.

2) Search space: In the context of HPO for pre-trained CNNs, the search space defines the range and type of hyperparameters that are optimized to improve model performance. This space encompasses various types of parameters that differ in structure and optimization complexity. The hyperparameter search space Λ is typically composed of categorical, discrete and continuous parameters.

Each category introduces different challenges for optimisation algorithms. Categorical parameters often require non-gradient-based search techniques, discrete parameters increase combinatorial complexity, and continuous parameters demand fine-grained exploration.

C. Metaheuristic Algorithms

Metaheuristics are population-based or single-solution optimisation methods that incorporate stochastic search strategies to efficiently explore large, complex landscapes. They are particularly suited to HPO, as they do not rely on gradient information and can handle mixed-variable and non-differentiable objective functions [26], [27].

Metaheuristics achieve success by balancing:

- Exploration: discovering new regions of the search space.
- Exploitation: refining promising candidate solutions [7], [8].
- Several algorithms have been applied to CNN HPO:
- Genetic Algorithms (GA): Inspired by natural evolution, used for optimising filter sizes and learning rates [28].
- Particle Swarm Optimisation (PSO): Mimics social behaviour of birds, effective in refining CNN configurations [29].
- Whale Optimisation Algorithm (WOA): Employs simulated hunting behaviour for global exploration [30].
- Harmony Search and Variants: Used for fine-tuning CNNs through adaptive exploration-exploitation trade-offs [31].

Among these, two algorithms stand out for their complementary strengths:

- Artificial Bee Colony (ABC): A swarm-based algorithm inspired by bee foraging behaviour [9]. ABC excels at exploration by maintaining population diversity through scouts and onlookers but suffers from weak local refinement in later iterations [11].
- Harris Hawks Optimisation (HHO): A predator-prey inspired method with dynamic exploration and exploitation phases [10]. HHO is effective in local

search but is prone to premature convergence in highly multimodal spaces [12].

Individual metaheuristic algorithms often underperform in the complex hyperparameter landscape of pre-trained CNNs, leading to stagnation, poor convergence, or lack of diversity [13]. To address these challenges, hybrid metaheuristics have emerged, aiming to combine complementary strengths while mitigating weaknesses [14]. Recent studies, such as PSO-GWO [32] and HHOS [33] highlight the effectiveness of hybridisation for deep learning optimisation.

Nevertheless, the integration of ABC and HHO for CNN hyperparameter tuning remains underexplored. This motivates the present study, which introduces I-HAHO, an integrative hybrid approach that leverages ABC's exploratory power and HHO's exploitative refinement under a diversity-based switching strategy.

III. Related Works

This section presents reviews on recent studies that apply hybrid metaheuristic algorithms to HPO in pre-trained CNN models. Recent studies have explored hybrid metaheuristic algorithms for HPO in pre-trained CNN models, yet consistent challenges remain across adaptability, accuracy, and computational efficiency dimensions.

Ulutas et al. [34] proposed a hybrid PSO-GWO approach for tuning hyperparameters within ensemble models for diabetes detection, achieving a peak accuracy of 98.10% using Random Forest classifiers. Their findings indicate that the hybrid method outperforms standalone Particle Swarm Optimization (PSO) and Grey Wolf Optimization (GWO), offering improved robustness and generalisation across classifiers.

Yan et al. [35] addressed key challenges in CNN hyperparameter tuning, including mixed-variable encoding, high computational cost, and convergence stability. They introduced GPPSO, a hybrid algorithm combining Gaussian Processes with PSO. The method incorporates a novel encoding scheme for mixed-variable hyperparameters and a surrogate model to reduce evaluation overhead. Experiments on CIFAR-10 and CIFAR-100 confirmed GPPSO's superior efficiency and accuracy, with practical relevance for applications such as metal fracture diagnosis.

Kumar and Kondaveeti [33] applied transfer learning with a Hybrid Hyperparameter Optimization Scheme (HHOS), integrating manual domain knowledge and random search strategies. Using pre-trained ImageNet models with additional trainable layers, they demonstrated that fine-tuning EfficientNetB0 achieved 99.12% accuracy, outperforming ResNet18 (93.24%). These results highlight the value of hybrid HPO in ecological classification tasks.

Akkuş et al. [36] explored HPO for CNN-based image classification, specifically diabetic retinopathy detection. Their study compared the Adolescent Identity Search Algorithm (AISA) with Bayesian Optimization (BO) across pre-trained models including AlexNet, MobileNetV2, ResNet18, and GoogLeNet. AISA-based tuning yielded higher accuracy with fewer iterations, demonstrating its efficiency in medical imaging contexts.

Kiliçarslan [32] investigated hybrid optimisation for cardiovascular disease prediction using 1D and 2D VGG-16 architectures. The study compared PSO, Cat Swarm Optimization (CSO), and hybrid PSO-GWO, showing that the hybrid consistently achieved the highest classification accuracy across medical and MNIST datasets.

Atteia et al. [37] further emphasised the importance of HPO in CNN-based diagnostic systems, applying optimisation techniques to diabetic maculopathy detection using optical coherence tomography and fundus retinography. Their work reinforces the critical role of HPO in enhancing model performance for clinical applications.

The summary of the related works is presented in Table II.

TABLE II. Summary of Related Works

Authors	Hybrid Method	Domain	Models / Datasets	Key Findings
Ulutaş et al. [34]	PSO + GWO	Diabetes detection (Ensemble Learning)	Random Forest, other classifiers	Achieved 98.10% accuracy; hybrid PSO-GWO outperformed standalone PSO and GWO, improving robustness and model performance.
Yan et al. [35]	GP + PSO (GPPSO)	CNN HPO, Metal fracture diagnosis	CNNs on CIFAR-10, CIFAR-100	Proposed mixed-variable encoding + surrogate model; improved efficiency, convergence, and accuracy compared to traditional methods.
Kumar & Kondaveeti [33]	Hybrid Hyperparameter Optimisation Scheme (HHOS) (manual + random search)	Ecological application(transfer learning)	EfficientNetB0, ResNet18 (pre-trained on ImageNet)	Fine-tuned EfficientNetB0 achieved 99.12%, ResNet18 achieved 93.24%; hybrid scheme improved accuracy and adaptability.
Akkuş et al. [36]	AISA + Bayesian Optimisation	Medical imaging (Diabetic Retinopathy detection)	AlexNet, MobileNetV2, ResNet18, GoogLeNet; Kaggle DR dataset	Hybrid AISA-based optimisation achieved higher accuracy with fewer iterations than BO; more efficient for CNN HPO in medical imaging.

IV. Materials and Methods

In this section, I-HAHO is introduced as our approach to improve the accuracy of the HPO for pre-trained CNN models.

A. Dataset

This study utilizes four benchmark datasets: CIFAR-10, CIFAR-100, SVHN, and TinyImageNet to evaluate the performance of the proposed hybrid metaheuristic HPO algorithms. These datasets were selected due to their widespread

adoption in deep learning literature and their varying levels of complexity, class granularity, and domain diversity, which support robust validation of adaptability and generalisation.

CIFAR-10 and CIFAR-100 were obtained from the official dataset repository maintained by the Canadian Institute for Advanced Research (CIFAR). Each dataset comprises 60,000 32×32 colour images. CIFAR-10 consists of 10 mutually exclusive classes, whereas CIFAR-100 is structured with 100 fine-grained categories across 20 superclasses. Both datasets were split into 50,000 training and 10,000 testing images. During preprocessing, images were normalized using dataset-specific mean and standard deviation values, converted to float tensors, and standardized to improve gradient stability during model training. No class rebalancing was performed, preserving original distribution characteristics for realism.

Street View House Numbers (SVHN), sourced from the official Stanford University repository, was employed in its cropped digits format, ensuring consistent spatial resolution with other datasets. Comprising 73,257 training and 26,032 test images, SVHN represents a real-world digit classification task under non-ideal lighting and alignment conditions. Raw .mat files were parsed and converted to tensor format. Images were resized to 32×32 pixels and normalized to zero mean and unit variance. Dataset labels were integer-encoded and subsequently transformed to one-hot format for compatibility with CNN classifiers. It has a real-world and noisy data characteristic which makes it a challenging test case.

TinyImageNet, a scaled-down variant of the ILSVRC ImageNet dataset, was downloaded from the official ImageNet challenge portal. It includes 200 object categories, with 100,000 training samples (500 images per class), 10,000 validation samples (50 per class), and 10,000 test samples. Unlike CIFAR and SVHN, the input resolution for TinyImageNet was preserved at 64×64 pixels to maintain semantic richness. Due to its relatively small sample size per class and high inter-class similarity, extensive preprocessing was applied, including data augmentation techniques such as random horizontal flipping, random cropping with padding, and brightness/contrast jittering to prevent overfitting and improve generalization under low-data regimes. It represents real-world scenarios and tests the scalability and efficiency of optimisation algorithms in complex tasks.

A summary of the datasets is provided in Table III, highlighting their complexity based on factors such as image resolution, number of classes, and dataset size.

TABLE III. SUMMARY OF DATASETS

Dataset	Number of Classes	Total Images	Image Resolution	Complexity
CIFAR-10	10	60,000	32x32	Moderate
CIFAR-100	100	60,000	32x32	High
TinyImageNet	1,000	1.2 million	Varies	Very High
SVHN	10	600,000+	32x32	High (Real-world)

B. Proposed I-HAHO Algorithm

The pseudocode for the I-HAHO is presented in the Algorithm 1.

Algorithm 1: I-HAHO

Input:
Population Size N, Maximum iterations T, Initial diversity threshold D_0,
Hyperparameters for ABC and HHO, Objective function $F(x)$
Output:
Best solution x_{best} with optimized hyperparameters.

Initialise the population X using Latin Hypercube Sampling
Set iteration counter $t = 0$
Calculate the initial diversity threshold $D_{\text{thresh}}(t) = D_0$
Evaluate the fitness $F(x)$ for each solution $x_i \in$ X
While $t <$ T:
    Calculate population diversity $D(t)$
    Update the diversity threshold $D_{\text{thresh}}(t)$
        Dynamic Phase-Switching Strategy:
            $if D(t) > D$thresh (High Diversity):
                Perform Modified ABC Exploration
            else $D(t) < D$thresh
                Perform Modified HHO Exploitation
            Evaluate the $F(x)$ of the updated solutions
            Update the global best solution x_{best} based on the fitness
            Calculate the change in fitness ΔF = |F_best(t) - F_best(t-1) |.
            Convergence Check:
            If ΔF < threshold or if $D(t) > D$min
            Increment t = t +1
End While
Return the best solution x_{best}

Fig. 4 illustrates the detailed operational flow of the proposed Integrative Hybrid Metaheuristic Algorithm (I-HAHO), which unifies exploration and exploitation through a diversity-based phase-switching mechanism. At each iteration, the population diversity $D(t)$ is computed and compared with a dynamic threshold $D_{\text{thresh}}(t)$. When $D(t) > D_{\text{thresh}}$, indicating sufficient population diversity, the algorithm prioritises Modified ABC exploration to expand the search space. This phase adaptively updates candidate positions, reinitialises stagnated solutions using Distance-Aware Novelty Reinitialisation (DANR), and invokes Chaos-Guided Candidate Recovery (CGCR) when diversity falls below acceptable limits, ensuring sustained global exploration. Conversely, when $D(t) < D_{\text{thresh}}$, the process transitions to Modified HHO exploitation, which refines promising regions through Refined Surround Prey, Escaping Energy, and Success-History-Based Adaptation (SHEA) to intensify local search and accelerate convergence.

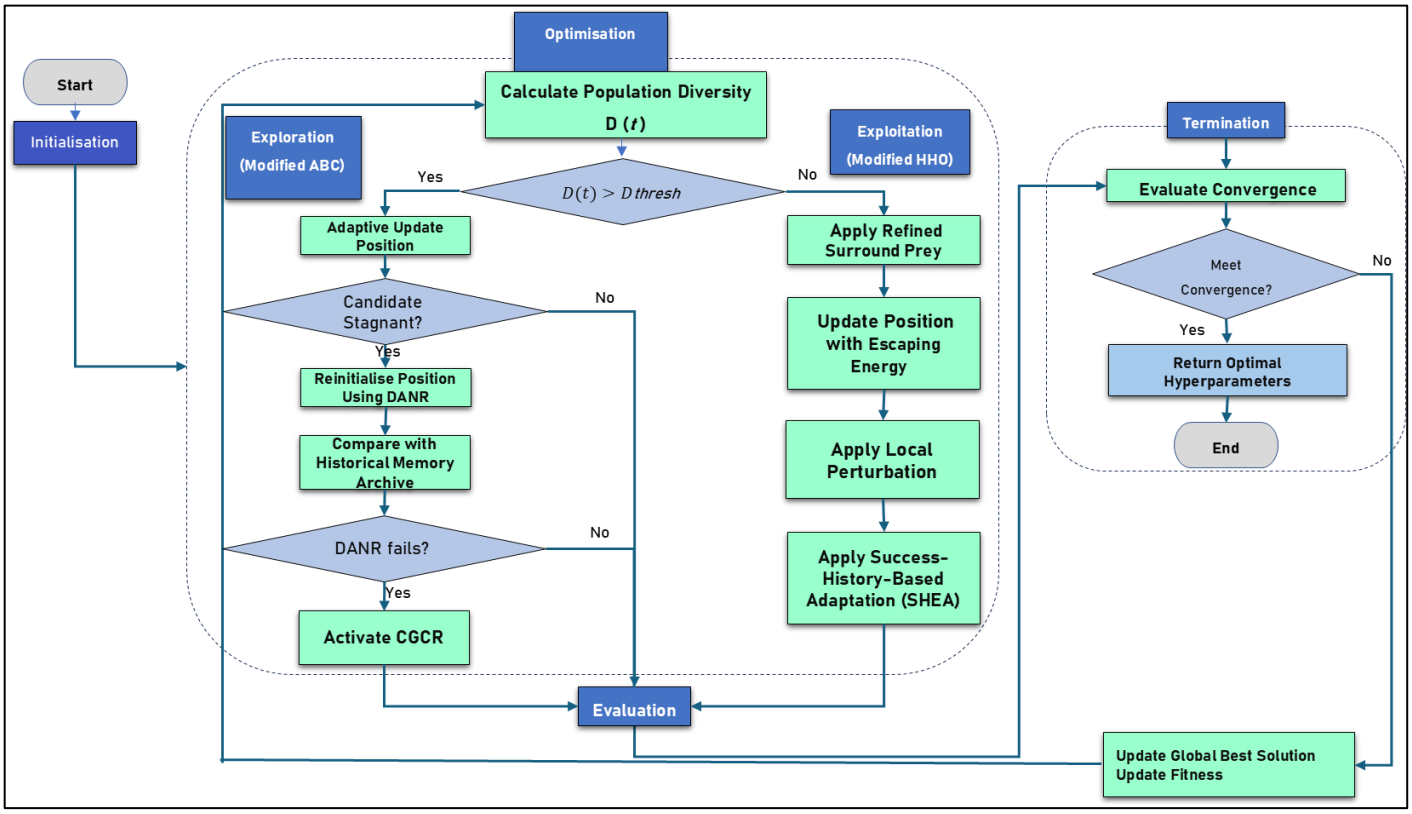


Fig. 4. Detailed algorithmic flowchart for I-HAHO with key strategies.

Unlike previous hybrid metaheuristics that rely on static schedules, random alternation, or fixed probability-based phase switching, the proposed diversity-driven framework continuously adapts based on the evolutionary state of the population. This allows I-HAHO to intelligently balance exploration and exploitation according to real-time feedback from the search dynamics, rather than predetermined iteration counts. The integration of DANR, CGCR, and SHEA further enhances the system's self-adaptive behaviour. DANR promotes novelty and prevents stagnation by introducing diversity-sensitive perturbations, CGCR leverages chaotic patterns to recover lost diversity while preserving promising regions, and SHEA enables adaptive parameter fine-tuning guided by historical success rates. Collectively, these mechanisms fill the gap left by earlier hybrid strategies that lacked dynamic responsiveness, allowing I-HAHO to navigate complex, multimodal hyperparameter landscapes more effectively and maintain balanced convergence performance across datasets and CNN architectures.

C. Performance Evaluation Measures

The evaluation measures employed in this research were selected to comprehensively assess the performance of the proposed I-HAHO algorithm.

1) Classification accuracy: Classification accuracy is the primary metric to evaluate the performance of HPO in pre-trained CNN models. It measures the proportion of correctly classified instances relative to the total instances.

$$\text{Accuracy}(\theta) = \frac{\text{Correctly Classified Instances}(\theta)}{\text{Total Instances}} \quad (1)$$

A higher classification accuracy indicates better optimisation of hyperparameters and superior adaptibility performance of the model.

V. Results and Discussion

A. Overall Comparative Performance

The experiment results for I-HAHO comparison with baseline HPO algorithms and canonical metaheuristic algorithms are presented in Table IV. The results demonstrate the highest accuracy achieved by different HPO algorithms across four pre-trained CNN models on various benchmark datasets. The accuracy results are presented in Table IV.

The proposed I-HAHO algorithm consistently outperformed all benchmarked HPO algorithms across the evaluated pre-trained CNN models and datasets. On the CIFAR-10 dataset, I-HAHO achieved superior validation accuracies of 93.194% (VGG16), 94.503% (ResNet50), and 93.008% (EfficientNet-B0), outperforming traditional and baseline search-based algorithms (GS, RS and BO). This consistent improvement illustrates I-HAHO's ability to effectively navigate the hyperparameter space through adaptive hybridization of global and local search strategies.

TABLE IV. ACCURACY RESULTS FOR ALL HPO ALGORITHMS AND I-HAHO

Model	Dataset	GS (%)	RS (%)	BO (%)	Vanilla ABC (%)	Vanilla HHO (%)	I-HAHO (%)
VGG16	CIFAR-10	88.2	88.6	89.2	87.5	87.7	**93.1**
	SVHN	95.3	95.6	96.2	94.7	95.0	**96.9**
	CIFAR-100	67.1	68.04	68.5	66.0	66.5	**78.5**
	Tiny ImageNet	57.8	58.3	58.9	57.0	57.4	**65.4**
ResNet50	CIFAR-10	90.5	90.9	91.3	89.0	89.3	**94.5**
	SVHN	93.0	93.7	94.1	92.5	92.9	**95.8**
	CIFAR-100	68.9	69.7	70.4	67.6	68.3	**76.0**
	Tiny ImageNet	61.2	61.9	62.5	60.2	60.9	**69.5**
EfficientNet-B0	CIFAR-10	89.0	89.7	90.1	87.5	87.9	**93.0**
	SVHN	87.3	87.9	88.4	86.0	86.5	**91.0**
	CIFAR-100	68.0	68.9	69.5	66.6	67.3	**75.5**
	Tiny ImageNet	61.8	62.4	62.9	59.6	60.3	**68.0**

On the more complex CIFAR-100 dataset, which includes 100 classes with fewer samples per class, I-HAHO achieved 78.507% (VGG16), 76.011% (ResNet50), and 75.512% (EfficientNet-B0), confirming its adaptability in high-dimensional, class-imbalanced learning tasks. For the SVHN dataset, characterized by large-scale digit recognition with significant intra-class variance and noise, I-HAHO achieved 96.903% (VGG16), 95.803% (ResNet50), and 91.007% (EfficientNet-B0). These results demonstrate I-HAHO's ability to maintain high accuracy and generalization performance even under challenging, real-world-like scenarios with substantial class overlap and varying input conditions.

On TinyImageNet, a particularly challenging dataset with 200 categories and lower resolution images, I-HAHO again demonstrated superior adaptability, achieving 65.497% (VGG16), 69.502% (ResNet50), and 68.013% (EfficientNet-B0). Notably, the EfficientNet-B0 model exhibited strong performance due to its compound scaling, which supports effective feature extraction even in low-resolution, high-category contexts. These outcomes suggest that I-HAHO not only improves training outcomes for well-known benchmarks but also scales effectively to more complex domains with less structured data distributions.

Collectively, these results are aligned with recent studies highlighting the limitations of single-strategy optimizers in deep learning [3], [37], [38] and emphasize the benefit of hybrid metaheuristics for pre-trained CNN models HPO tasks.

VI. CONCLUSION

In conclusion, the proposed Integrative Hybrid Artificial Bee Colony–Harris Hawks Optimization (I-HAHO) algorithm improves hyperparameter optimization (HPO) for pre-trained convolutional neural networks (CNNs) through an adaptive hybrid structure. The framework integrates a modified Artificial Bee Colony (ABC) algorithm for global exploration with a modified Harris Hawks Optimization (HHO) algorithm for local exploitation. This integration provides a balanced search process that minimizes premature convergence and sustains population diversity, which conventional approaches such as grid search, random search, and Bayesian optimization fail to achieve. The adaptive phase-switching mechanism regulates exploration and exploitation dynamically according to population diversity and fitness variance, allowing the optimiser to adjust its search intensity based on the topological characteristics of the loss landscape. This adaptivity enables I-HAHO to sustain population diversity during early iterations and progressively intensify local refinement as convergence nears, resulting in consistent and efficient optimisation in high-dimensional, non-convex search spaces. Experimental results on CIFAR10, CIFAR100, SVHN, and TinyImageNet datasets across VGG16, ResNet50, and EfficientNetB0 architectures demonstrate up to 6.9% improvement in classification accuracy compared to baseline optimizers. These results validate IHAHO as a scalable and generalizable hyperparameter optimization framework that improves search efficiency, convergence stability, and model adaptability across diverse deep learning tasks. The diversity-based phase-switching mechanism enables dynamic control over exploration and exploitation, supporting robust performance in high-dimensional, nonconvex spaces.

REFERENCES

[1] K. Wagh, K. S. Suresh, M. Sayed, and S. Shahane, "Exploring transfer learning in image analysis using feature extraction with pre-trained models," *ICTACT Journal on Image & Video Processing*, vol. 14, no. 3, 2024.

[2] R. N. Wanjiku, L. Nderu, and M. Kimwele, "Improved transfer learning using textural features conflation and dynamically fine-tuned layers," *PeerJ Computer Science*, vol. 9, p. e1601, 2023.

[3] S. Iqbal, A. Qureshi, A. Ullah, J. Li, and T. Mahmood, "Improving the robustness and quality of biomedical CNN models through adaptive hyperparameter tuning," *Applied Sciences*, vol. 12, no. 22, p. 11870, 2022.

[4] F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., *Automated Machine Learning: Methods, Systems, Challenges*. Cham: Springer, 2019.

[5] V. Silva, "Optimizing transfer learning and fine-tuning hyperparameters in image classification problems with firefly algorithm," in *Proc. Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2024)*, pp. 484–495, 2024. doi: 10.5753/eniac.2024.245032.

[6] S. B. Şüyun, M. Yurdakul, Ş. Taşdemir, and S. Biliş, "Triple-stream deep feature selection with metaheuristic optimization and machine learning for multi-stage hypertensive retinopathy diagnosis," *Applied Sciences*, vol. 15, no. 12, p. 6485, 2025.

[7] A. Gaspar, D. Oliva, E. Cuevas, D. Zaldívar, M. Pérez, and G. Pajares, "Hyperparameter optimization in a convolutional neural network using metaheuristic algorithms," in *Metaheuristics in Machine Learning:*

Theory and Applications. Cham: Springer International Publishing, pp. 37–59, 2021.

[8] R. Narayanan and N. Ganesh, "A comprehensive review of metaheuristics for hyperparameter optimization in machine learning," in *Metaheuristics for Machine Learning: Algorithms and Applications*, pp. 37–72, 2024.

[9] D. Karaboga and B. Basturk, "Artificial bee colony algorithm," *Journal of Global Optimization*, vol. 39, no. 3, pp. 459–471, 2007.

[10] 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.

[11] B. Akay and D. Karaboga, "A modified artificial bee colony algorithm for real-parameter optimization," *Information Sciences*, vol. 192, pp. 120–142, 2012.

[12] T. Yang, J. Fang, C. Jia, Z. Liu, and Y. Liu, "An improved Harris hawks optimization algorithm based on chaotic sequence and opposite elite learning mechanism," *PLoS One*, vol. 18, no. 2, p. e0281636, 2023.

[13] S. Sharma, P. Singh, and P. Kaur, "Metaheuristic optimization algorithms for hyperparameter tuning in deep learning: A comprehensive review and future directions," *Artificial Intelligence Review*, vol. 55, pp. 4895–4953, 2022. doi: 10.1007/s10462-021-10032-7.

[14] A. Seyyedabbasi, "A hybrid multi-strategy optimization metaheuristic algorithm for multi-level thresholding color image segmentation," *Applied Sciences*, vol. 15, no. 13, p. 7255, 2025.

[15] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[16] K. Simonyan and A. Zisserman, "Very deep convnets for large-scale image recognition," *Computing Research Repository*, 2014.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning," *Image Recognition*, vol. 7, no. 4, pp. 327–336, 2015.

[18] L. Mahin Sheikh, A. Shaikh, A. Sandupatla, R. Pudale, A. Bakare, and M. Chavan, "Classification of simple CNN model and ResNet50," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, 2024. doi: 10.22214/ijraset.2024.60677.

[19] Q. Ji, J. Huang, W. He, and Y. Sun, "Optimized deep convolutional neural networks for identification of macular diseases from optical coherence tomography images," *Algorithms*, vol. 12, no. 3, p. 51, 2019.

[20] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. International Conference on Machine Learning (ICML)*, pp. 6105–6114, May 2019.

[21] E. Şafak and N. Barışçı, "Detection of fake face images using lightweight convolutional neural networks with stacking ensemble learning method," *PeerJ Computer Science*, vol. 10, p. e2103, 2024.

[22] M. A. K. Raiaan, S. Sakib, N. M. Fahad, A. Al Mamun, M. A. Rahman, S. Shatabda, and M. S. H. Mukta, "A systematic review of hyperparameter optimization techniques in convolutional neural networks," *Decision Analytics Journal*, vol. 11, p. 100470, 2024.

[23] E. Tuba, N. Bačanin, I. Strumberger, M. Tuba, and R. Jovanović, "Convolutional neural networks hyperparameters tuning," in *Artificial Intelligence: Theory and Applications*. Cham: Springer, pp. 65–84, 2021. doi: 10.1007/978-3-030-72711-6_4.

[24] M. Feurer and F. Hutter, "Hyperparameter optimization," *Automated Machine Learning*, vol. 3, no. 2, pp. 33–50, 2019.

[25] A. Morales-Hernández, I. Van Nieuwenhuyse, and S. Rojas Gonzalez, "A survey on multi-objective hyperparameter optimization algorithms for machine learning," *Artificial Intelligence Review*, vol. 56, pp. 8043–8093, 2023. doi: 10.1007/s10462-022-10359-2.

[26] F. Peres and M. Castelli, "Combinatorial optimization problems and metaheuristics: Review, challenges, design, and development," *Applied Sciences*, vol. 11, no. 14, p. 6449, 2021.

[27] S. Adamu, H. Alhussian, N. Aziz, S. J. Abdulkadir, A. Alwadin, A. A. Imam, A. Garba, and Y. Saidu, "Optimizing hyperparameters for improved melanoma classification using metaheuristic algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 10, pp. 531–540, 2023. doi: 10.14569/IJACSA.2023.0141062.

[28] M. R. Apriyadi and D. P. Rini, "Hyperparameter optimization of support vector regression algorithm using metaheuristic algorithm for student performance prediction," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 2, 2023.

[29] N. Bacanin, T. Bezdan, E. Tuba, I. Štrumberger, and M. Tuba, "Optimizing convolutional neural network hyperparameters by enhanced swarm intelligence metaheuristics," *Algorithms*, vol. 13, no. 3, p. 67, 2020. doi: 10.3390/a13030067.

[30] A. Brodzicki, "The whale optimization algorithm approach for deep learning hyperparameter tuning," *Sensors*, vol. 21, no. 23, p. 8003, 2021. doi: 10.3390/s21238003.

[31] D. Liu, H. Ouyang, S. Li, and C. Zhang, "Hyperparameters optimization of convolutional neural network based on local autonomous competition harmony search algorithm," *Journal of Computational Design and Engineering*, vol. 10, no. 4, pp. 1280–1297, 2023.

[32] S. Kiliçarslan, "PSO+GWO: A hybrid particle swarm optimization and grey wolf optimization based algorithm for fine-tuning hyper-parameters of convolutional neural networks for cardiovascular disease detection," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 1, pp. 87–97, 2023.

[33] S. V. S. Kumar and H. K. Kondaveeti, "Bird species recognition using transfer learning with a hybrid hyperparameter optimization scheme (HHOS)," *Ecological Informatics*, 2024. doi: 10.1016/j.ecoinf.2024.102510.

[34] H. Ulutas, R. B. Günay, and M. E. Sahin, "Detecting diabetes in an ensemble model using a unique PSO-GWO hybrid approach to hyperparameter optimization," *Neural Computing and Applications*, vol. 36, no. 29, pp. 18313–18341, 2024.

[35] H. Yan, C. Zhong, Y. Wu, L. Zhang, and W. Lu, "A hybrid-model optimization algorithm based on the Gaussian process and particle swarm optimization for mixed-variable CNN hyperparameter automatic search," *Frontiers of Information Technology & Electronic Engineering*, vol. 24, no. 11, pp. 1557–1573, 2023.

[36] E. Akkuş, U. Bal, F. Ö. Koçoğlu, and S. Beyhan, "Hyperparameter optimization of pre-trained convolutional neural networks using adolescent identity search algorithm," *Neural Computing and Applications*, vol. 36, no. 4, pp. 1523–1537, 2024.

[37] G. Atteia, N. Samee, E. El-kenawy, and A. Ibrahim, "CNN-hyperparameter optimization for diabetic maculopathy diagnosis in optical coherence tomography and fundus retinography," *Mathematics*, vol. 10, no. 18, p. 3274, 2022.

[38] A. Amou, H. Costa, and R. Ferreira, "A novel MRI diagnosis method for brain tumor classification based on CNN and Bayesian optimization," *Healthcare*, vol. 10, no. 3, p. 494, 2022.

A Novel Performance-Based Time Series Forecast Combination Method and Applications with Neural Networks

M. Burak Erturan
General Directorate of State Hydraulic Works, Directorate of Region 13, Antalya, Türkiye

***Abstract*—Performance-based forecast combination approaches determine the weights of the individual forecasts based on the inverse average error for a past time interval. However, although the performances are calculated for a time span, the aim is mostly a one-step-ahead time-point forecast. In these classical methods, a relatively higher prediction error of a single past time-point spreads and decreases the performance value of the model, even though the model is highly successful on other time-points in the interval. In this study, a novel approach is presented where performance of each past time-point prediction is calculated separately. Instead of taking the inverse average error for a pre-determined past time interval, prediction performance is calculated for each past data point separately using the normalized inverse absolute error, then the average performances are calculated for past time interval to get the combination weights. To be able to measure the performance of the presented methodology, it is applied on three well-known time series data. Seven different models of neural networks, based on multi-layer perceptron and extreme learning machines are used to model, forecast and form the combination forecasts. Moreover, four different performance-based combination techniques, two central tendency-based benchmark combination methods and the naïve model are employed for comparison. The obtained results show that proposed methodology is a powerful and robust technique and superior to all performance-based combination techniques compared.**

Keywords*—Combination forecast; performance-based combination; neural networks; multi-layer perceptron; extreme learning machine

I. Introduction

Time series forecasting has become a major research area in academics. Researchers have developed various methodologies to improve the accuracy of forecasting. From classical statistical models to machine learning [22] and heuristic models, many models have been developed. Moreover, researches showed that using more than one forecast at the same time often gives better results than a single forecast [1]. This development has led the researchers to hybridize or combine different forecasts and resulted with hybrid and combination forecasting models.

Combination forecast is done by combining multiple forecasts from different models and is usually better than a single forecast. The interested real world time series data is probably cannot be explained by a specific model, because of the consisting different processes like time-varying trends, changing seasonality and structural breaks [1]. Moreover, selecting the "best model" is somewhat problematic because of the sample, parameter and model uncertainties. Sample uncertainty is the problem that different sample sizes result different models and various parameter estimates. Parameter estimation uncertainty arises from different algorithms and various model setups. Furthermore, different model structures result with various parameter restrictions, which may cause for example the estimation problem [2]. Under these circumstances, combining multiple forecasts has been a highly preferred methodology.

The main question of the forecast combination is that, "how to combine the individual forecasts". Different answers to this question result various combination methods. An important class of combination procedures is performance-based forecast combination. That class of methods determine the weights of individual forecasts according to corresponding models' past performances. Some models use inverse past accuracy measures of the models like mean squared error (MSE), root mean squared error (RMSE) or symmetric mean absolute percentage error (sMAPE) to calculate the weights, such that the forecast of the model with a low error measure has a high weight in the forecasting pool. Another idea is, instead of using inverse error measures directly, ranking the models according to their past accuracy and employing the inverse of these rank values as weights, which is called rank-based forecast combination.

There are some characteristics of current performance-based combination methodologies arise from the mean error metrics they are based on. In inverse error based combination methods, average of the errors for some past time span is used. In these methods, the averaging procedure gives general information about the time span performance, but not enough information about the performances for past time-points individually. For example, a possible high error value of only one past prediction might cause a low performance; even other past predictions are highly accurate. On the other hand, rank-based model assigns the weights through the inverse sequence of numbers. Regardless of how much the variation between the models' past accuracies, the weights are pre-determined and constant. Therefore, rank-based combinations limit the weights to only a set of possible values [3].

Forecasting techniques are (including forecast combinations) used mostly for one-step ahead forecasting. Even when the forecast target consists of multi-steps ahead time-points, usually, forecasting is done step by step. It means

that, first the next step is forecasted, and then the subsequent forecast is done using previous forecast as if it is the real value. Therefore, pointwise performance values of the forecasting models are highly important. Classical performance-based combination techniques take an average error value of the individual models for a past time-span (for the whole past training data-span, or a more recent one) and use it to assign the performance values. In this paper, a novel performance-based combination methodology is proposed, which is different from classical performance-based methods. First, the performances of each past time-point prediction of individual models are calculated separately by inverse absolute error (IAE). Then, these point performance values are normalized by min-max normalization within other models for each past time point data; such that, the best performance value is one, the worst is zero and the others are in between. The min-max normalization among different models assigns a relative performance value and holds equal performance scales ([0, 1]). Moreover, since recent data holds more information and is more significant, the weights are calculated by averaging the time-point performances of recent data by the rolling window method which is often used in the literature [1].

To test the performance of the proposed model, three well-known time series data are used for modeling and forecasting purposes. These datasets are Canadian lynx, Wolf's sunspot and GBP/USD foreign exchange series. Seven individual neural network models are constructed for combination. Three of the neural network models are based on multi-layer perceptron (MLP) and four of them are different sub-models of extreme learning machines (ELM). Obtained results from proposed combination technique are compared with the results of seven individual models, four other performance-based models, equally weighted simple mean/ median combinations and the naïve model.

The rest of the paper is organized as follows: In Section II, a review of literature related to current is presented. Section III describes all methodologies used in the research. First, MLP and ELM models are explained, which are used for individual forecasts. Then, performance-based combination models are described, which are used for comparison purposes. After that, proposed forecast combination model is introduced and explained. Lastly, the details of the data, models and the software used for application purposes are presented. Section IV presents the application results and findings. Section V concludes with discussions of the study's implications, limitations, and future work.

II. Related Work

Since the work of Bates and Granger [4], forecast combination literature has been grown substantially. From simple combination procedures to much more complex systems, researchers have developed many models and methods. However, forecast combination studies show that simpler combination methods are very successful and hard to beat [1]. The most popular one is taking the simple arithmetic mean of the individual forecasts, ignoring the past information of the models. This procedure is quite successful, robust and easy to perform [3]. Other simple and robust methodologies related to central tendency (CT) are taking the median or the mode of the forecasting pool. In contrast to the simplicity of the mean and median, mode combination needs numerous forecasts or some techniques like discretizing the data or kernel density estimation [5]. Therefore, the simple mean and the median combinations are the most referenced CT combinations, and there is still debate among the scholars whether the mean or the median is more successful [1]. Other than simple equally weighted average model, some researchers applied the trimmed or winsorized mean for combination forecasts. Results show that especially when there are high variations among the forecasts, trimmed and winsorized mean methods produce good results [5]. Simple combination rules are very successful and easy to implement. Therefore, among other and more sophisticated methods, simple combination methods are the choice of many researchers as a benchmark for testing new combination methodologies. There are other more complex methods than simple combinations, such as linear combinations, nonlinear combinations and combining by learning. A comprehensive classification of combination methods can be found in study [1].

Linear combination methods assign the weights of individual forecasts according to their accuracy in a linear combination. Optimal weights approach aims to minimize the variance by optimizing the weights [6], while in linear regression based methods this optimization is done by regression [7]. In criteria based techniques, weights of the individual forecasts are determined by using information criteria, such as Akaike's (AIC) or Bayesian (BIC) information criteria [8, 9].

Performance-based forecast combinations are in the class of linear combination methods, where each individual forecast has a weight according to its performance on past data. Differences in performance measurement are resulted with different models in the literature. In study [10], IRMSE (inverse root mean squared error) model is used with different combination models to forecast day-ahead spot electricity prices. It is found that, while IRMSE method is one of the best performer models, there is no single model dominating the others for all datasets. In the work of [11] different forecasting models are combined with weights determined by their sMAPE values. It is determined that the selection of the forecasting models is crucial for the combination forecast to be successful. Another performance-based model is rank-based model used in [12]. Rank based models determine the weights of individual models according to inverse of their rankings. This model is also used by study [13] for tourism demand forecasting and compared with other combination methods. Performance-based combinations are used to forecast various time series, such as livestock prices [14], oil prices [15], ozone concentration, and airline passengers [16].

III. Methodology

A. Multilayer Perceptron

Artificial Neural Networks (ANN) are computational models inspired by the human brain's neural networks. An ANN is composed of interconnected nodes (neurons) organized in layers; including input, hidden, and output layers. They mimic the way real neurons communicate; and are particularly useful in pattern recognition, classification, optimization, and

time series prediction. ANNs can model complex nonlinear relationships and handle large datasets with numerous variables.

A multi-layer perceptron (MLP) is the most commonly used type of artificial neural network that consists of multiple layers of nodes; including an input layer, one or more hidden layers, and an output layer. The network uses a feedforward path for data processing and backpropagation for training and optimization. Training MLPs involves challenges such as entrapment in local minima, convergence speed, and sensitivity to initialization. MLPs have strong generalization capabilities but are prone to overfitting, where the network becomes too dependent on the training data. Techniques such as regularization and cross-validation are often employed to mitigate overfitting. MLPs are widely used for time series modeling. As a mathematical expression for time series y_t, the one-step ahead forecast $\hat{y}_{(t+1)}$ of the real data y_{t+1} can be calculated as [5]:

$$\hat{y}_{t+1} = \beta_0 + \sum_{l=1}^{L} \beta_l g\left(b_i + \sum_{i=1}^{I} w_{li} x_i\right) \quad (1)$$

In Eq. (1), I is the number of the inputs x_i, and L is the number of hidden nodes. β and w are the weights of the output and hidden layers respectively, where $\beta=[\beta_1, \beta_2, \ldots, \beta_L]$ and $w=[w_1, w_2, \ldots, w_L]$. β_0 and b_i are the biases acting like intercept in the regression process. Lastly, $g(.)$ is the transfer function and is usually hyperbolic tangent or logistics function.

The backpropagation algorithm is a widely used method for training MLPs by adjusting the weights of the connections to minimize the error in predictions. There are three stages of backpropagation, in the first stage the calculations are done forward through the final output by using activation function. Then, obtained error is propagated backwards, starting from the output layer. Finally, the weights are changed in order to minimize errors. These stages continue until certain conditions are reached. Backpropagation is effective but can be slow and may be stuck in local minima [17]. Another training algorithm is resilient backpropagation (RPROP), which is an advanced training algorithm designed to improve the efficiency and performance of MLPs by addressing some of the limitations of traditional backpropagation methods. RPROP is an advanced version of backpropagation that adjusts the weight updates based on the sign of the gradient rather than its magnitude, leading to faster convergence and improved performance [17, 18]. As an improvement of RPROP, a weight backtracking mechanism can be added. Weight backtracking retracts a previous update for some or all weights, while whether taking back a step or not is decided by heuristics [19].

B. Extreme Learning Machine

Extreme learning machine (ELM) is a single layer feedforward network (SLFN). Different from other neural networks, the weights of connections between input and hidden layers are selected randomly and not trained further. Therefore, instead of training all the weights in the network, only the output weights (connecting hidden and output layer) are calculated. The output weights can be optimized with the approaches like least squares or equivalents. ELM has fast learning rate with strong generalization performance. Moreover it shows high efficiency in training, particularly when data is limited [20, 21]. A representation of ELM architecture is in Fig. 1.

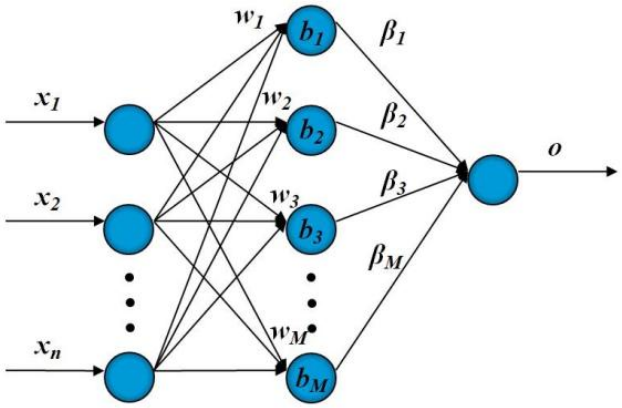


Fig. 1. Architecture of an extreme learning machine network with n input, M hidden, and 1 output nodes.

Suppose we have N arbitrary distinct training samples (x_j, t_j), each containing of n inputs and m outputs where $j = 1, 2, \ldots, N$. Therefore, $x_j = [x_{j1}, x_{j2}, \ldots, x_{jn}]^T \in Rn$ is the input vector and $tj = [tj1, tj2, \ldots, t_{jm}]^T \in$ Rm is the output vector. The output (o_j) of the standard SLFN model with M hidden nodes and the activation function of $g(.)$ can be modeled as [20]:

$$o_j = \sum_{i=1}^{M} \beta_i g\left(w_i x_j + b_i\right) \quad (2)$$

In Eq. (2), $w_i = [w_{i1}, w_{i2}, \ldots, w_{in}]^T$ is the weight vector connecting inputs to the ith hidden node. $\beta_i = [\beta_{i1}, \beta_{i2}, \ldots, \beta_{in}]^T$ the weight vector connecting ith hidden node to outputs, and bi is the bias (threshold) of ith hidden node. The standard SLFN with M hidden nodes with activation function $g(.)$ can approximate these N samples with zero error means that:

$$\sum_{i=1}^{M} \beta_i g\left(w_i x_j + b_i\right) = t_j \quad (3)$$

where $j = 1, 2, \ldots, N$. These N equations can be written in a more compact form as:

$$\boldsymbol{H\beta = T} \quad (4)$$

where,

$$\boldsymbol{H} = \begin{bmatrix} g(w_1 x_1 + b_1) & \ldots & g(w_M x_1 + b_M) \\ g(w_1 x_2 + b_1) & \ldots & g(w_M x_2 + b_M) \\ \vdots & \ddots & \vdots \\ g(w_1 x_N + b_1) & \ldots & g(w_M x_N + b_M) \end{bmatrix}_{NXM} \quad (5)$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \beta_2^T \\ \vdots \\ \beta_n^T \end{bmatrix}_{Mxm} \quad (6)$$

$$\boldsymbol{T} = \begin{bmatrix} t_1^T \\ t_2^T \\ \vdots \\ t_N^T \end{bmatrix}_{Nxm} \quad (7)$$

After setting the input weights and biases of the hidden layer, the output weight vector β can be obtained by a series of linear equations transformations. If the number M of hidden

nodes is equal to the number N of distinct training samples, $M = N$, matrix H is square and invertible. Because the input weight vectors wi and the hidden biases b_i are randomly chosen, the SLFN can approximate these training samples with zero error. However, in most cases the number of hidden nodes is much less than the number of distinct training samples, such that $M << N$. Then, H is a not a square matrix and there may not exist w_i, b_i, β_i ($i = 1, 2, ..., M$) such that $\boldsymbol{H}\beta = \boldsymbol{T}$. Then, the smallest norm least squares solution of the above linear system is [20]:

$$\hat{\beta} = \boldsymbol{H}^{\dagger}\boldsymbol{T} \tag{8}$$

where, $\boldsymbol{H}^{\dagger}$ represents the generalized (Moore-Penrose) inverse matrix of the output.

Different from traditional neural networks, parameters between hidden and output layer in the ELM structure can be obtained by minimizing least squares solutions with the provided training data [23, 24, 25]. Therefore, using ELM to obtain the output weights is consisted of three steps. First, randomly selecting numerical values between zero and one to set input weights and the bias of the hidden layer. Second, calculation of the output matrix $\boldsymbol{H}$. Finally, calculation of the output weights β. On the other hand, there are some possible problems with ELM networks. First, ELM can effectively approximate to any complex systems with huge numbers of hidden nodes, but in this case, the generalization might decrease. Another is that, the least square method cannot provide good estimates for weights when there exist many outliers in data. To overcome these deficiencies, researchers presents some developments like regularization techniques. Some main methods for this purpose are lasso, ridge and stepwise regression models.

Generally, in regression models a common target is keeping the model as simple as possible. Regularization models like LASSO (Least absolute shrinkage and selection operator) and Ridge regressions punish the complex model by adding a penalty term to the cost function. Ridge regression adds squared magnitude of the coefficients as penalty term, while LASSO adds absolute value of the coefficients as penalty term to the cost function. LASSO shrinks the less important variable coefficients to zero and causes a less complex structure [21]. Moreover, stepwise regression uses forward adding and backward deleting method to the variables to find the best fitting combination of independent variables for prediction [26].

C. Performance-Based Forecast Combinations

Performance-based forecast combination is a method used to improve the accuracy of predictions by linearly combining multiple individual forecasts by assigning weights to each forecast based on their past performances. This combination procedure ignores the correlations among individual forecasts. Moreover, the success of these models reconfirms that models ignoring correlations are more successful than other models, because correlations are poorly estimated in practice and should be ignored in weight calculations [1, 3].

One common branch of performance-based methodologies is inverse error technique, which uses inverse error values of the models directly as weights. Commonly used error metrics are mean squared error (MSE), root mean squared error (RMSE) and symmetrical mean absolute percentage error (sMAPE). In inverse error model, the weight of the ith model among N models with $j=[1,2,...N]$ is:

$$w_i = \frac{\varepsilon_i^{-1}}{\sum_{j=1}^{N}\varepsilon_i^{-1}} \tag{9}$$

where, ε is the error value of the model and might be MSE, RMSE or sMAPE according to which error metrics the combination is based on. Then the combined forecast for time t with N individual model and individual forecasts $\hat{y}_{tj}$ is:

$$\hat{y}_t = \sum_{j=1}^{N}\hat{y}_{tj}w_j \tag{10}$$

Note that, Eq. (9) is valid under the assumption of independence of errors. In the case of error-dependence (i.e. nested models), one of the models encompasses the other and the optimal combining weights are trivially either zero or one [27].

Another performance-based technique is combination forecasting with inverse rank based weights, which involves integrating multiple individual forecasts into a single forecast by assigning weights to each forecast based on their past performance with a ranking system. Proposed by [12], in the rank-based methodology, the weights of the individual forecasts are determined by the inverse of the ranks of the models, which are identified according to their past performance (usually MSE). The weight of the ith model among N models with $j=[1,2,...N]$ is:

$$w_i = \frac{Rank_i^{-1}}{\sum_{j=1}^{N}Rank_i^{-1}} \tag{11}$$

where, $Rank_i=[1,2,...,N]$ is the rank of the ith model's forecast according to model's past performance (MSE).

D. Proposed Combination Methodology

Performance determination with inverse error metrics uses an average value of performance for a pre-determined past time period. It is impossible to determine a point by point performance from these mean (average) error values. For example, a very bad performance with a high error value of a model in a past single point effects the whole performance no matter how successful the model in other past data points. On the other hand, rank-based methods not just use the mean error values, but also assign the performances of the models with a pre-determined values; which ignores the real performance differences between the competitive models. Furthermore, most of the forecasting researches are interested in one-step ahead forecasting for a multi-step period, including this study. Therefore, measuring the performance of each past time point separately can give more reliable values if forecasting procedure is one-step ahead.

Presented methodology takes past prediction performances into account separately for each time-point. First, inverse of absolute error (IAE) values for each past point predictions are calculated. Then, for each past data point, competitive models' IAE values are normalized within. A min-max normalization is applied such that the best prediction takes the value of one, the worst prediction is zero and the others are in between. The min-max normalization provides an equal performance scale

([0, 1]) for each past prediction. Obtained values are the performances of the models on past data points. Finally, the weights of the models are calculated by averaging the past performances. Proposed methodology takes neither the whole training set values nor just the values of test set into account. A rolling window method is used to take into account of the recent values whether they are training or test data. In this paper, a rolling window of 10 past predictions is used. Mathematical notation of the presented weight procedure is presented in Eq. (12).

$$p_{i,t} = NORM\left(AE_{i,t}^{-1}\right) = \frac{AE_{i,t}^{-1} - \min\left(AE_t^{-1}\right)}{\max\left(AE_t^{-1}\right) - \min\left(AE_t^{-1}\right)} \quad (12)$$

In Eq. (12), $p_{i,t}$ represents the performance value of the ith candidate model for time t. *NORM(.)* is the min-max normalization function, where AE^{-1} is the inverse absolute error. Then, the weights and final combination forecast for time t is:

$$w_i = \frac{\sum_{j=t-1}^{t-Z} p_{i,j}}{Z} \quad (13)$$

$$\hat{y}_t = \sum_{j=1}^{N} \hat{y}_{tj} w_j \quad (14)$$

where, Z is the number of past data to be used for weight wi of the ith model forecast, N is the number of the candidate models with individual forecasts of ytj, and j=[1,2,…N].

The usage of the proposed combination method can be summarized in three steps:

- The absolute error values for each past predictions of an individual model are calculated separately. The inverse of the AEs are normalized (min-max normalization) within other models for the same time points to get the time point performances of the models.
- For each model, time-point performances for Z past prediction are averaged for pre-determined time period to find the weights of the individual models.
- Combination forecast is found using individual forecasts and related weights.

The workflow of the proposed combination model used in this study is presented in Fig. 2.

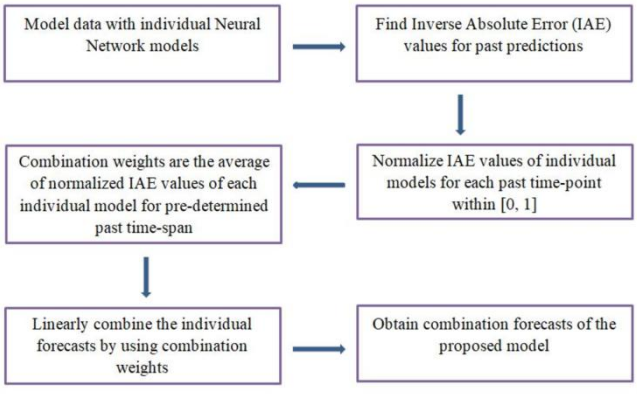


Fig. 2. The workflow of the proposed forecast combination model used in this study.

Proposed methodology has same characteristics with other performance-based models. It ignores the correlations among individual models like other performance-based models. It is a linear combination method using weights of the individual models, which are calculated by models' past performances. Proposed model's difference from other performance-based models is the calculation procedure of the past performances.

E. Data and Application

In this study, three well-known data sets in time series literature are used for the application purposes. They are Canadian lynx, Wolf's sunspot and British Pound/United States Dollar (GBP/USD) exchange rate data sets. All three datasets have various characteristics, and are used often in time series literature as benchmark datasets [28].

Canadian lynx dataset shows the number of yearly trapped lynx around North Canada Mackenzie River between 1821 and 1938. It has total 114 data points and is an important dataset in time series literature with nonlinear characteristics [29]. Wolf's sunspot data, known with nonlinear behavior, contains yearly sunspot numbers in the period of 1700-1987 and has 288 data points. Sunspot data is identified as nonlinear and non-Gaussian and used to test nonlinear models for their performances [28]. Lastly, GBP/USD dataset includes weekly values between 1821 and 1934, and is composed of 731 data points. Exchange rate, and in general, financial time series forecasting is a very difficult task. Various models are developed but few are successful to beat a random walk (naïve) model [29].

For the sake of consistency with the related literature [28-31] the logarithmic transformation with base 10 (log10) of Canadian lynx, and natural logarithmic transformation (ln) of GBP/USD exchange rate data sets are used in the application stage. Sunspot and transformed lynx datasets are stationary, while transformed exchange rate time series is nonstationary. As can be seen from the related literature [28-31], if a conventional time series model like ARIMA (Autoregressive integrated moving averages) to be used, this dataset should be integrated (differenced) with order 1 to be stationary. However, as neural networks can model nonstationary data directly, natural logarithmic transformation of the exchange rate data is used directly to be consistent with the aforementioned literature. On the other hand, successful modeling of nonstationary data is very difficult with respect to stationary data for all conventional and advanced time series models.

In Fig. 3 to 5, graphical representation of the Canadian lynx, Wolf's sunspot and GBP/USD datasets are presented respectively. In Fig. 3 and Fig. 5 the real data and log transformed data used could be seen separately. In Table I, descriptive statistics of the datasets are presented.

Seven different neural network models are used for modelling and forecasting purposes. These models are used for proposed combination forecast methodology and for comparison. Three of the models are in MLP class: MLP with classical backpropagation (MLP-BP), MLP with resilient backpropagation (MLP-RP) and MLP with resilient backpropagation and weight backtracking (MLP-RPB). Other four models are various ELM models: ELM with LASSO

regression (ELM-L), ELM with ridge regression (ELM-R), ELM with stepwise regression (ELM-S) and ELM with classical linear regression (ELM-LN). Other than these seven models, six different combination models are used for comparison purposes. Two of them are benchmark combination models: Classical equally weighted simple mean (C-MEAN), and median (C-MEDIAN) models. Last four models are other performance-based combination models: Combination models with inverse MSE (C-MSE), RMSE (C-RMSE) and sMAPE (C-SMAPE) based weights, and rank-based combination model (C-RANK). In addition, the naïve model (N) is added to the pool of the models for comparison.

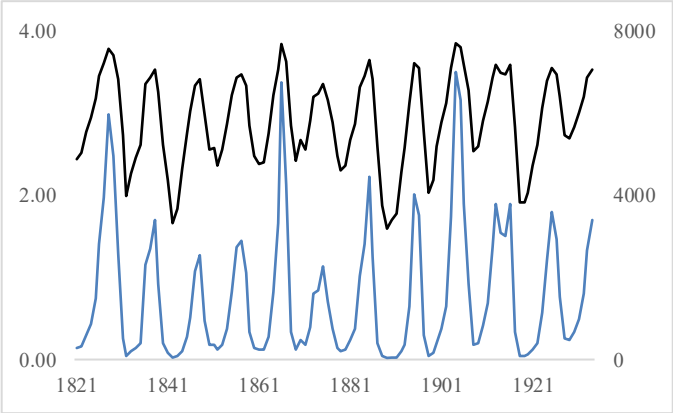


Fig. 3. Canadian lynx dataset. Real values (blue, right axis) and logarithmic transformed (log10) values (black, left axis).

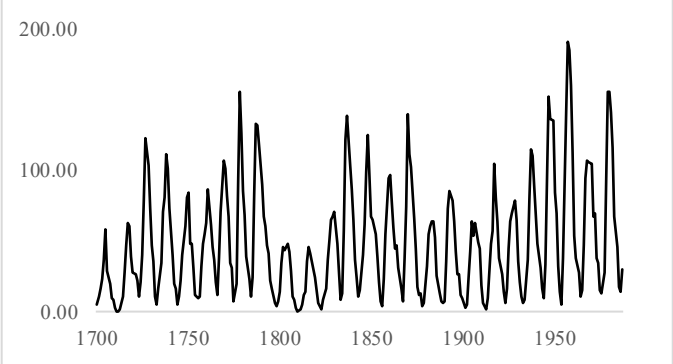


Fig. 4. Wolf's sunspot dataset.

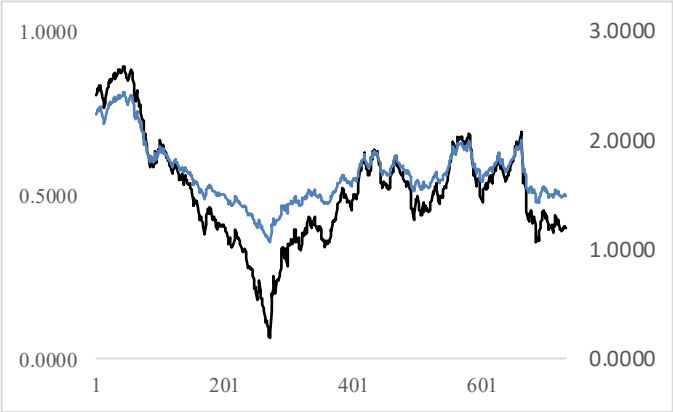


Fig. 5. GBP/USD exchange rate dataset. Real values (blue, right axis) and natural logarithmic transformed (ln) values (black, left axis).

TABLE I. DESCRIPTIVE STATISTICS OF THE DATASETS

	Canadian lynx (Log 10)	Wolf's sunspot	GBP/USD (Ln)
Data Number	114	288	731
Min	1.59	0.00	0.0628
Max	3.84	190.20	0.8909
Median	2.89	39.00	0.5043
Mean	2.90	48.43	0.5089
Std. Dev.	0.56	39.36	0.1616

Software used for this study is R Statistical Software version 4.4.1 [32]. Package nnfor [33] is used mainly with neuralnet [34] package. nnfor package is used for time series modeling and forecasting with neural networks. All seven neural networks are three layered with a single hidden layer and all activation functions are hyperbolic tangent (tanh). Since the datasets are well known and employed many times by other researchers, network architectures and training-test data compositions are determined to be compatible with the literature. MLP model architecture is 7-5-1 for lynx dataset, 4-4-1 for sunspot dataset, and 7-6-1 for GBP/USD data [28, 29, 31, 35, 36]. An exception is the number of hidden nodes in ELM networks. The nnfor package assigns the number of hidden neurons for ELM models automatically. ELMs start with a very large hidden layer (100 nodes), and then prune it as much as needed [33]. Therefore, number of the input nodes are 7, 4 and 7 for lynx, sunspot and exchange rate data sets respectively. Number of the hidden nodes are assigned by the nnfor package. Each individual neural network model is retrained 20 times and combined with mean values for a more robust forecast. Moreover, all forecasting procedures are one step forecasting. Forecast performances of the models are compared using MSE, MAD (Mean absolute deviation) and MAPE (Mean absolute percentage error) error metrics.

IV. RESULTS

Canadian lynx data is composed of 114 data points. First 100 data points are used as training set, while last 14 data points are test set to be forecasted. The architecture of all MLP models are 7-5-1. On the other hand, ELM-L/ELM-R models are both 7-91-1 and ELM-S/ELM-LN model architectures are 7-40-1. Obtained results are tabulated in Table II.

Results show that proposed model and C-RMSE model are superior to all other performance-based models. In terms of MAD and MAPE, presented model has the best results among other performance-based models. However, equally weighted mean and median combinations are more successful than proposed and other performance-based combinations. Among the individual models, ELM with LASSO regression model gives the best results. But still, in terms of MSE, C-MEAN combination model is superior to all individual neural network models. An important finding is that, all individual and combination models are better than the naïve model.

Wolf's sunspot dataset includes 288 data points with 221 of the data is allocated as training set, while last 67 data is formed as test set. MLP model architectures are 4-4-1, ELM-L/ELM-R

models are 4-100-1 and ELM-S/ELM-LN models are 4-40-1. Performance results of the models are in Table III.

TABLE II. FORECASTING RESULTS OF CANADIAN LYNX DATASET

Forecasting Model		MSE (x10^{-2})	MAD (x10^{-1})	MAPE
Individual Models	MLP-BP	2.5993	1.2961	4.48%
	MLP-RP	2.5859	1.2501	4.31%
	MLP-RPB	2.5366	1.2963	4.49%
	ELM-L	2.3318	1.2739	4.27%
	ELM-R	4.5764	1.9828	6.44%
	ELM-S	2.5889	1.4073	4.81%
	ELM-LN	2.9211	1.4732	5.06%
CT Based Combinations	C-MEAN	2.2895	1.2918	4.37%
	C-MEDIAN	2.3900	1.2892	4.42%
Performance Based Combinations	C-MSE	2.4439	1.3430	4.59%
	C-RMSE	2.3740	1.3250	4.51%
	C-RANK	2.4741	1.4131	4.63%
	C-SMAPE	2.3921	1.3266	4.53%
	Proposed Model	**2.3907**	**1.3214**	**4.51%**
	Naïve	6.8734	2.3088	7.77%

TABLE III. FORECASTING RESULTS OF WOLF'S SUNSPOT DATASET

Forecasting Model		MSE (x10^{2})	MAD (x10^{1})	MAPE
Individual Models	MLP-BP	2.9186	1.3089	35.44%
	MLP-RP	2.8160	1.2843	33.60%
	MLP-RPB	2.8085	1.2909	34.59%
	ELM-L	3.2518	1.4155	40.03%
	ELM-R	3.5500	1.4817	42.90%
	ELM-S	2.7776	1.2724	34.55%
	ELM-LN	2.7846	1.2925	35.45%
CT Based Combinations	C-MEAN	2.7681	1.2935	35.57%
	C-MEDIAN	2.8138	1.2958	35.28%
Performance Based Combinations	C-MSE	2.7594	1.2888	35.35%
	C-RMSE	2.7629	1.2910	35.46%
	C-RANK	2.7832	1.2903	35.22%
	C-SMAPE	2.7689	1.2933	35.52%
	Proposed Model	**2.7353**	**1.2836**	**35.22%**
	Naïve	9.2073	2.2964	54.84%

Results presented in Table II shows that proposed combination model is the best model among all other combination models. On the other hand, in terms of MAD and MAPE error metrics, two different individual models show best performances. MLP-RP and ELM-S models seems to be the best models among individual models. Also, it can be seen that, performance-based combination methods have slightly better performances than central tendency based combinations. Naïve model has the worst performance among all competent models with significantly high error metrics.

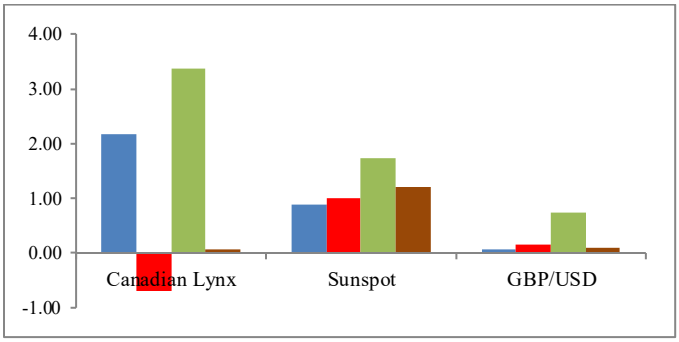
731 data points of GBP/USD exchange is divided to two as training and test sets, such that 679 data points are training and 52 data points are test data. MLP models are in 7-6-1 architecture, whereas, ELM-L/ELM-R models are 7-100-1, and ELM-S/ELM-LN models are 7-40-1. Obtained results are in Table IV.

TABLE IV. FORECASTING RESULTS OF GBP/USD DATASET

Forecasting Model		MSE (x10^{-4})	MAD (x10^{-2})	MAPE
Individual Models	MLP-BP	1.9434	1.1342	2.82%
	MLP-RP	1.9025	1.1216	2.79%
	MLP-RPB	1.8855	1.1028	2.74%
	ELM-L	1.8297	1.0972	2.72%
	ELM-R	1.8300	1.0973	2.72%
	ELM-S	1.8563	1.1144	2.77%
	ELM-LN	1.9003	1.1261	2.79%
CT Based Combinations	C-MEAN	1.8341	1.1020	2.74%
	C-MEDIAN	1.8391	1.1029	2.74%
Performance Based Combinations	C-MSE	1.8315	1.0995	2.73%
	C-RMSE	1.8329	1.1008	2.73%
	C-RANK	1.8436	1.0998	2.73%
	C-SMAPE	1.8321	1.1004	2.73%
	Proposed Model	**1.8301**	**1.0990**	**2.73%**
	Naïve	1.8298	1.1010	2.73%

Results show that almost all individual models and combinations show high performances. All combination models and the naïve model show very close performance values in all error metrics. Best individual models are ELM-L and ELM-R models, and they are slightly better than the forecast combinations. On the other hand, among all combination models, proposed combination methodology shows the best performance. Another important result drawing attention is the successful performance of the naïve model with respect to its less accurate results on sunspot and lynx datasets.

In Fig. 6, the percentage improvement by using the proposed model instead of other performance-based models in terms of error metrics is presented.



(a)

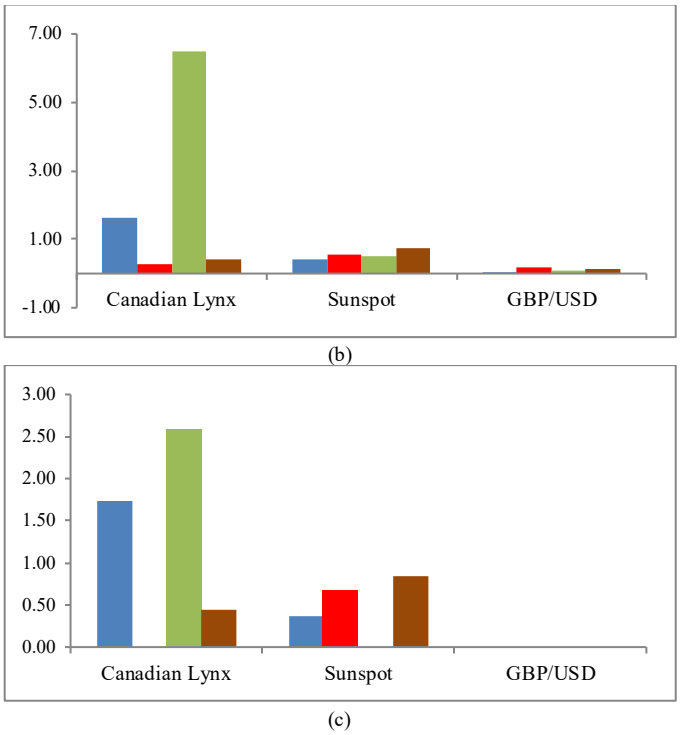


Fig. 6. Percentage improvement of the proposed model in terms of Mean Squared Error (a), Mean Absolute Deviation (b) and Mean Absolute Percentage Error (c) from other performance-based models (C-MSE (blue), C-RMSE (red), C-RANK (green) and C-SMAPE (brown).

V. Conclusion

In this study, a novel performance-based forecast combination methodology is presented. The novelty of the method is originated from the calculation of the individual models' past performances. Classical performance-based combination methods use a mean error value (MSE, MAD, sMAPE etc.) for a pre-determined past time-period. Conversely, proposed methodology measures the performance of each past time point separately, and then takes the average of these time point performances for the interested past time period. Furthermore, after computation of inverse absolute error values for each time point, a min-max normalization is applied within other individual models. That procedure causes all models' time-point performance values to be between zero and one, while the best model's is one and the worst model's becomes zero. After that, each model's performances are calculated by averaging the time point performances along the pre-determined time-period. To evaluate the performance of the proposed model, three well-known time series are used. Seven different neural network models based on MLP and ELM are trained. Moreover, simple mean and median combinations, four main performance-based combinations, and naïve model are formed for comparison purposes. Comparisons are done with MSE, MAD and MAPE error metrics.

Obtained forecasting results show that proposed combination methodology is superior to other performance-based combination methods in almost all occasions. Additionally, proposed technique is superior to central tendency based benchmark models of equally weighted mean and median combinations in sunspot and exchange rate datasets. In lynx dataset, mean and median combinations mostly show better performances. Proposed technique increases the forecasting accuracy more than other compared performance-based combination models. It beats the naïve model in lynx and sunspot dataset, whereas the naïve model is slightly better in GBP/USD time series in terms of MSE.

Considering the obtained results, it can be stated that the proposed forecast combination methodology is a very successful and robust performance-based technique. Moreover, findings show that calculating the performance value of an individual model for each past data point separately causes more accurate combination weights than computing directly from an average error value for that past time-span. This contribution to the forecast model performance metrics can be tested and used in further researches. Application to different time series data on various subjects and using different individual time series models can help the generalization of the model. Furthermore, studying with various past time periods for performance evaluation and comparing with different combination models are some other possible research extents. On the other hand, proposed model's performance evaluation is necessarily more time consuming than conventional performance-based methods. However, especially under the circumstances where recent data is used as presented study is, it is tolerable.

References

[1] X. Wang, R. J. Hyndman, F. Li, and Y. Kang, "Forecast combinations: An over 50-year review," Int. J. Forecast, vol. 39, no. 4, pp. 1518-1547, Oct.–Dec. 2023, doi: 10.1016/j.ijforecast.2022.11.005.

[2] N. Kourentzes, D. Barrow, and F. Petropoulos, "Another look at forecast selection and combination: Evidence from forecast pooling," Int. J. Prod. Econ., vol. 209, pp. 226-235, Mar. 2019, doi: 10.1016/j.ijpe.2018.05.019.

[3] V. Genre, G. Kenny, A. Meyle, and A. Timmermann, "Combining expert forecasts: Can anything beat the simple average?," Int. J. Forecast., vol. 29, pp. 108-121, Jan.-Mar. 2013, doi: 10.1016/j.ijforecast.2012.06.004.

[4] J.M. Bates, and C. W. J. Granger, "The combination of forecasts," Oper. Res. Q., vol. 20, no. 4, pp. 451-468, Dec. 1969, doi: 10.1057/jors.1969.103.

[5] N. Kourentzes, D. K. Barrow, and S. F. Crone, "Neural network ensemble operators for time series forecasting," Expert Syst. Appl., vol. 41, pp. 4235-4244, Jul. 2014, doi: 10.1016/j.eswa.2013.12.011.

[6] A. Timmermann, "Forecast combinations," in Handbook of economic forecasting, vol. I, G. Elliott, C. W. J. Granger, and A. Timmermann, Eds. Elsevier, 2006, pp. 135-196, doi: 10.1016/S1574-0706(05)01004-9.

[7] C. Conflitti, C. De Mol, and D. Giannone, "Optimal combination of survey forecasts," Int. J. Forecast., vol. 31, no. 4, pp. 1096-1103, Oct.-Dec. 2015, doi: 10.1016/j.ijforecast.2015.03.009.

[8] S. Kolassa, "Combining exponential smoothing forecasts using Akaike weights," Int. J. Forecast., vol. 27, no. 2, pp. 238-251, Apr.-Jun. 2011, doi: 10.1016/j.ijforecast.2010.04.006.

[9] F. Petropoulos, N. Kourentzes, K. Nikolopoulos, and E. Siemsen, "Judgmental selection of forecasting models," J. Oper. Manag., vol. 60, pp. 34-46, Jun. 2018, doi: 10.1016/j.jom.2018.05.005.

[10] J. Nowotarski, E. Raviv, S. Trück, and R. Weron, "An empirical comparison of alternative schemes for combining electricity spot price forecasts," Energy Econ., vol. 46, pp. 395-412, Nov. 2014, doi: 10.1016/j.eneco.2014.07.014.

[11] M. Pawlikowski and A. Chorowska, "Weighted ensemble of statistical models," Int. J. Forecast., vol. 36 no. 1, pp. 93-97, Jan.-Mar. 2020, doi: 10.1016/j.ijforecast.2019.03.019.

[12] M. Aiolfi and A. Timmermann, "Persistence in forecasting performance and conditional combination strategies," J. Econom., vol. 135, no. 1, pp. 31-53, Nov.-Dec. 2006, doi: 10.1016/j.jeconom.2005.07.015.

[13] R. R. Andrawis, A. F. Atiya, and H. El-Shishiny, "Combination of long term and short term forecasts, with application to tourism demand forecasting," Int. J. Forecast., vol. 27, no. 3, pp. 870-886, Jul.-Sep. 2011, doi: 10.1016/j.ijforecast.2010.05.019.

[14] L. Ling, D. Zhang, A. W. Mugera, S. Chen, and Q. Xia, "A Forecast Combination Framework with Multi-Time Scale for Livestock Products' Price Forecasting," Math. Probl. Eng., Article ID 8096206, Oct. 2019, doi: 10.1155/2019/8096206.

[15] C. Baumeister and L. Kilian, "Forecasting the real price of oil in a changing world: A forecast combination approach," J. Bus. Econ. Stat., vol. 33, no. 3, pp. 338-351, Jul. 2015, doi: 10.1080/07350015.2014.949342.

[16] J. P. Donate, P. Cortez, G. G. Sanchez, and A. S. De Miguel, "Time series forecasting using a weighted cross-validation evolutionary artificial neural network ensemble," Neurocomputing, vol. 109, pp. 27-32, Jun. 2013, doi: 10.1016/j.neucom.2012.02.053.

[17] W. Saputra, A. P. Windarto, and A. Wanto, "Analysis of the Resilient Method in Training and Accuracy in the Backpropagation Method," Int. J. Inform. Comput. Sci., vol. 5, no. 1, pp. 52-56, Mar. 2021, doi: 10.30865/ijics.v5i1.2922

[18] M. Riedmiller and H. Braun, "A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm," Proc. IEEE Int. Conf. Neural Netw., vol. 1, pp. 586-591, Mar.-Apr. 1993, doi: 10.1109/ICNN.1993.298623

[19] C. Igel and M. Hüsken, "Empirical evaluation of the improved Rprop learning algorithms," Neurocomputing, vol. 50, pp. 105-123, Jan. 2003, doi: 10.1016/S0925-2312(01)00700-7

[20] G. B. Huang, Q. Y. Zhu, and C. K. Siew, "Extreme learning machine: Theory and applications," Neurocomputing, vol. 70, no. 1-3, Dec. 2006, doi: 10.1016/j.neucom.2005.12.126

[21] D. T. Várkonyi and K. Buza, "Extreme learning machines with regularization for the classification of gene expression data," in Proceedings of the 19th Conference Information Technologies-Applications and Theory (ITAT 2019), B. Petra, H. Martin, H. Tomáš, P. Matúš, and R. Rudolf, Eds., CEUR Workshop Proceedings, 2019, pp. 99-103. https://ceur-ws.org/Vol-2473/paper11.pdf.

[22] S. Ding, X. Xu, and R. Nie, "Extreme learning machine and its applications," Neural Comput. Appl., vol. 25, pp. 549-556, Sep. 2014, doi: 10.1007/s00521-013-1522-8.

[23] G. Huang, G. B. Huang, S. Song, and K. You, "Trends in extreme learning machines: a review," Neural Netw., vol. 61, pp. 32-48, Jan. 2015, doi: 10.1016/j.neunet.2014.10.001.

[24] J. Wang, S. Lu, S. H. Wang, and Y. D. Zhang, "A review on extreme learning machine," Multimed. Tools Appl., vol. 81, pp. 41611-41660, Dec. 2022, doi: 10.1007/s11042-021-11007-7.

[25] Y. Yang, H. Zho, Y. Gao, J. Wu, Y. G. Wang, and L. Fu, "Robust penalized extreme learning machine regression with applications in wind speed forecasting," Neural Comput. Appl., vol. 34, pp. 391-407, Jan. 2022, doi: 10.1007/s00521-021-06370-3.

[26] A. Alqahtani, "Engineering the Energy Gap of Cupric Oxide Nanomaterial Using Extreme Learning Machine and Stepwise Regression Algorithms," J. Nanomater., Article ID 4797686, Oct. 2021, doi: 10.1155/2021/4797686.

[27] T. E. Clark and M. W. McCracken, "Combining Forecasts from Nested Models," Oxford B. Econ. Stat., vol. 71, no. 3, pp. 303-329, Apr. 2009, doi: 10.1111/j.1468-0084.2009.00547.x.

[28] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," Neurocomputing, vol. 50, pp. 159-175, Jan. 2003, doi: 10.1016/S0925-2312(01)00702-0.

[29] M. Khashei and M. Bijari, "Hybridization of the probabilistic neural networks with feed-forward neural networks for forecasting," Eng. Appl. Artif. Intell., vol. 25, pp. 1277-1288, Sep. 2012, doi: 10.1016/j.engappai.2012.01.019.

[30] C. N. Babu and B. E. Reddy, "A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data," Appl. Soft Comput., vol. 23, pp. 27-38, Oct. 2014, doi: 10.1016/j.asoc.2014.05.028.

[31] A. Atesongun and M. A. Gulsen, "Hybrid Forecasting Structure Based on Arima and Artificial Neural Network Models," Appl Sci., vol. 14, Article ID 7122, Aug. 2024, doi: 10.3390/app14167122.

[32] R Core Team, "R: A language and environment for statistical computing v. 4.4.1 [software]," R Foundation for Statistical Computing, 2020, https://www.R-project.org.

[33] N. Kourentzes, "nnfor: Time Series Forecasting with Neural Networks v. 0.9.9 [software package]," R, Nov. 2023, doi: 10.32614/CRAN.package.nnfor.

[34] S. Fritsch, F. Guenther, M. N. Wright, M. Suling, and S.M. Mueller, "neuralnet: Training of Neural Networks v. 1.44.2 [software package]," R, Feb. 2019, doi: 10.32614/CRAN.package.neuralnet.

[35] Z. Hajirahimi and M. Khashei, "Sequence in Hybridization of Statistical and Intelligent Models in Time Series Forecasting," Neural Process. Lett., vol. 54, Oct. 2022, doi: 10.1007/s11063-020-10294-9

[36] R. Adhikari and R. K. Agrawal, "A Homogeneous Ensemble of Artificial Neural Networks for Time Series Forecasting," Int. J. Comput. Appl., vol. 32, no. 7, pp. 1-8, Oct. 2011, doi: 10.48550/arXiv.1302.6210

Integration of Color QR-Code Technology in Biometric Data Encoding and Facial Identity Systems

Nazym Kaziyeva[1], Kalybek Maulenov[2]*, Ruslan Ospanov[3], Abzhan Khamza Mukhtaruly[4]
Department of Information Security, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan[1, 3, 4]
Department of Software Engineering, Kostanay Regional University named after Akhmet Baytursinuly, Kostanay, Kazakhstan [2]

***Abstract*—This paper presents an enhanced algorithm for the generation of color biometric QR codes capable of encoding facial image data, anthropometric parameters, and personal identity information simultaneously within a single RGB-based QR structure. The proposed approach extends existing monochrome QR models by integrating optimized image decomposition, modular QR block generation, and multi-channel RGB encoding to achieve higher data density, improved privacy protection, and better readability under various lighting and compression conditions. The algorithm was implemented in Python using the OpenCV library, ensuring compatibility with contemporary biometric systems, embedded devices, and mobile platforms. Experimental evaluations conducted on standard face databases demonstrate the method's robustness in terms of decoding accuracy, distortion resilience, and information integrity. Furthermore, the study explores new applications such as animated QR codes and photo–sketch hybrid datasets for training and validation purposes. The results highlight the potential of color biometric QR technology for secure identification, access control, and digital identity verification, offering a novel bridge between computer vision and information security.**

Keywords*—Color biometric QR code; facial image encoding; RGB channel decomposition; biometric data integration; secure identification; facial recognition; QR animation; identity encoding; privacy protection; data capacity; OpenCV; computer vision

I. Introduction

The integration of biometric technologies with modern data encoding methods is becoming increasingly vital in domains such as identity verification, access control, healthcare, and border security. Among data encoding methods, QR (Quick Response) codes have gained widespread popularity due to their fast readability, error correction capabilities, and versatility in encoding alphanumeric and binary information. Traditionally used for logistics, marketing, and product tracking, QR codes are now being investigated as compact carriers of sensitive biometric information.

Conventional QR codes, however, are typically monochromatic and limited in their data capacity. Moreover, they lack the ability to incorporate rich biometric con-tent such as facial imagery or structural anthropometric data in an integrated and se-cure manner. This limitation has prompted researchers to explore multi-channel encoding strategies that utilize color QR codes to store and transmit biometric data.

Recent studies have proposed the use of RGB-layered QR codes where different information types (e.g., facial images, anthropometric coordinates, and personal data) are embedded in separate color channels. These "color biometric QR codes" introduce a new paradigm by combining visual and data-driven representations into a single graphical object, potentially enhancing security, capacity, and usability in biometric systems. In the works conceptual, mathematical models, methods of creating color biometric QR codes (Color BIO QR codes) were first presented; these software algorithms were developed in MATLAB. As is known, MATLAB is used for scientific developments; to adapt algorithms for modern systems today it is necessary to use neural networks.

This paper presents an advanced and extensible method for generating color biometric QR codes using open-source tools and modern computer vision techniques. The proposed system allows the encryption of facial images, structured anthropometric data, and identity metadata into a single color QR code, using RGB decomposition and modular QR code generation. Different color QR structures are further investigated (such as PIA, PIP, and PIS), implement a prototype system using Python and OpenCV, and evaluate the generated codes on standard face datasets.

- The contributions of this work include:
- A novel algorithm for multi-layer biometric QR code generation;
- An implementation framework suitable for real-world applications and mobile environments;
- A set of experimental validations demonstrating readability, robustness, and data density;
- A discussion on potential applications including identity documents, biometric test bases, and photo-sketch encoding.

This study provides a significant step toward the practical adoption of color QR codes in biometric systems and opens new directions for research in secure and com-pact data encoding.

II. Related Work

Biometric systems have evolved significantly in the past two decades, particularly in the area of facial recognition, where algorithms now achieve near-human accuracy under controlled conditions. With the growing demand for portability, efficiency, and integration in real-world systems, researchers have explored compact methods to store and transmit biometric data — including through barcode technologies.

Corresponding author.

QR codes, introduced in the 1990s by DENSO WAVE [1] and standardized in ISO/IEC 18004 [2], have been widely used in various sectors due to their high data capacity, fast readability, and support for error correction. However, traditional black-and-white QR codes are limited in their ability to encode large volumes of complex data, such as facial images or anthropometric feature sets [3]. Their primary use in biometrics has been confined to storing short strings such as user IDs, cryptographic hashes, or database references rather than biometric patterns themselves.

Early work by Soldek et al. [4] and Heeter [5] explored the use of biometric markers in secure transaction systems. Previously, in works [6-12], proposed options for providing biometric data directly in QR structures. Papers in 2018 and 2019 introduced the concept of color biometric QR codes, where different types of information — such as grayscale facial photos (Photo), identity metadata (INFO), and anthropometric coordinates (ANTRO) — are encryption separately in the red, green, and blue channels of an RGB image. Their work laid the groundwork for using QR codes as composite carriers of personal identity data and the name color biometric QR codes were pro-posed for the new object.

Further developments include sketch-to-photo recognition studies (e.g., Wang and Tang [13]), which emphasize the importance of creating multimodal biometric datasets for testing. These studies show that including sketch-based representations in biometric systems can improve flexibility in law enforcement or forensic applications. Some researchers have also investigated animated QR codes as a novel way to represent dynamic or evolving biometric content. In study [14], it was proposed that the use of animation-based QR codes for facial biometrics, allowing different biometric features to be encoded across time-sequenced frames. Other works explored the use of animated QR formats in digital access passes and multimodal visualization [15]. Additionally, recent optical approaches to QR-based biometric encoding, such as the XOR-based color encoding system [16], demonstrate further potential for secure and high-capacity biometric data representation.

Despite these advances, several gaps remain. First, the security of stored data in colored QR codes is often not considered — encryption and access control mechanisms remain understudied. Second, empirical validation of these QR structures on real biometric datasets is still limited, especially with respect to scale. Finally, most prior studies lack open-source implementations that combine facial imagery with structured biometric data in a unified and extensible framework — limiting reproducibility and wider adoption in practical systems.

This work aims to address these gaps by presenting an open-source implementation of a color biometric QR code generator using Python, and by evaluating its performance across multiple biometric data types. Our study builds upon and extends prior research in this field by providing a unified framework for encoding visual and structured biometric data within a single QR code image, along with experimental validation and practical application scenarios.

III. Proposed Method

A. System Architecture Overview

The proposed system aims to generate a color biometric QR code that simultaneously encodes three types of data: (1) a facial image, (2) textual identity metadata, and (3) structured biometric features (e.g., anthropometric coordinates). The overall architecture consists of the following functional blocks (Fig. 1):

- Facial Image Analyzer: Performs preprocessing, normalization, and feature extraction from the input facial image;
- Message Preparation Block: Formats textual and biometric data into standardized messages suitable for QR encoding;
- QR Code Generator: Generates separate grayscale QR matrices for the INFO and ANTRO data;
- RGB Decomposition Unit: Splits the facial image into red, green, and blue channels;
- QR Code Composer: Merges the grayscale QR matrices and selected image components into a single RGB image representing the color QR code.
- The inputs to the system are:
- A color facial image I of size M × N × 3;
- Metadata INFO (e.g., full name, ID number);
- Biometric data ANTRO (e.g., coordinates of facial landmarks).

The output is a final color QR code matrix QRxxx, where each of the RGB layers carries different encoded content.

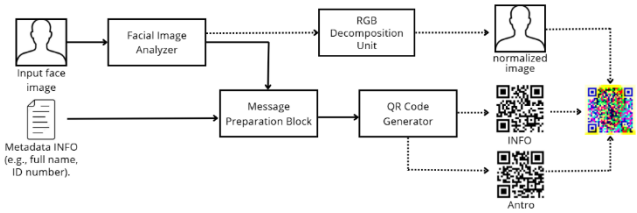


Fig. 1. System architecture of the proposed color biometric QR code generator.

B. Preprocessing of Facial Image

The input to the system is a color facial image I of arbitrary dimensions M×N×3. In order to integrate this image into the RGB structure of a color QR code, it must be resized, normalized, and prepared in a format compatible with standard QR matrix dimensions. Additionally, geometric consistency is important for biometric analysis and fusion with associated QR layers.

QR codes have strictly square dimensions defined by the version parameter V, which determines the number of modules in the QR matrix:

$$M_{QR} = 17 + 4V,\ V \in [1,40] \quad (1)$$

Let $I_{orig} \in R^{M \times N \times R}$ be the input image. We define the normalized image $I_{norm} \in R^{Mqr \times Nqr \times 3}$ as:

$$I_{norm} = resize(square(I_{orig}), M_{QR}, M_{QR}) \quad (2)$$

where resize() uses bicubic interpolation, and square() denotes symmetric cropping or padding.

If landmark-based alignment is available (e.g., using Dlib or Mediapipe), a similarity transformation is applied to align key facial features (eyes, nose base) horizontally. This ensures consistency in appearance across different samples, which is important if the RGB channels will later be compared or decoded separately.

Let $L = \{(x_i, y_i)\}_{i=1}^{n}$ be the set of facial landmarks. The image can be transformed using a similarity matrix $T \in R^{2 \times 3}$ to produce the aligned version:

$$I_{aligned}(x,y) = I_{norm}(T^{-1}(x, y)) \quad (3)$$

This step is optional but recommended if Anthropometric QR code encoding is involved.

To improve visual quality after integration into QR structures, Contrast enhancement techniques may be applied such as:

- Histogram equalization (per channel or grayscale),
- CLAHE (Contrast Limited Adaptive Histogram Equalization),
- Gamma correction.

The enhanced image is then converted to 8-bit unsigned integer format (uint8), as re-quired by most QR image fusion routines:

$$I_{final} = uint8(255 * normalize(I_{enhanced})), \quad (4)$$

The final preprocessed image $I_{final} \in R^{M_{QR} \times M_{QR} \times 3}$ is passed to:

- RGB decomposition (see Section 3.4), and;
- Landmark extraction (if ANTRO data is needed).
- This ensures that all downstream QR generation steps operate on a consistent spatial layout and image resolution.

C. Generation of Component QR Codes

- To embed textual metadata and biometric features into separate channels of the color QR code, we generate two grayscale QR matrices:
- QR_INFO: encodes textual identity information;
- QR_ANTRO: encodes structured anthropometric features extracted from the facial image.

These QR codes are generated independently using standardized QR encoding libraries and then fused into the RGB structure in later steps.

The identity metadata typically includes a person's name, ID number, birthdate, or other text-based identification data. This information is first compiled into a plain-text or JSON structure. Example format:

{ "id": "245891203", "name": "John Smith","dob": "1990-05-21"}, (5)

This structured message is encoded into a QR code using Python libraries such as:

- qrcode (pure Python, easy to use),
- segno (supports full QR versions and error correction),
- PyQRCode (compact and efficient).

Anthropometric data represents biometric landmarks — key points on the face used for recognition (e.g., eyes, nose tip, mouth corners). These points are usually extracted using facial landmark detectors such as:

- Dlib (68-point shape predictor);
- Mediapipe face mesh (468 points);
- OpenCV Haar Cascades + post-processing.
- The coordinates are encoded as a JSON or CSV string, e.g:

{ "landmarks": [[34,45],[67,45],[50,60],...]}, (6)

The QR code contains structured biometric data and can be decoded later by splitting the color QR image and reading the appropriate channel.

To ensure reliable encoding and decoding of biometric and identity data, the generation of component QR codes must follow certain design principles. The size of each QR code is determined by the version number V, which defines the matrix dimension according to the formula MQR=17+4V. Typically, versions between 10 and 20 are suitable, as they provide enough space to encode structured data while maintaining manageable image size.

For maximum robustness, especially when the QR code is later embedded into a color image, the highest level of error correction (Level H) is recommended. This al-lows recovery of up to 30% of the QR code content in the presence of noise, distortion, or compression.

The QR encoding should be performed in byte mode to support structured for-mats like JSON, ensuring compatibility with a wide range of decoding tools. Libraries such as qrcode or segno in Python allow fine control over version, error correction level, and matrix resolution.

To facilitate seamless integration with the RGB facial image, the generated QR codes must be resized to exactly match the dimensions of the normalized face image. For this purpose, the box size parameter in the QR library should be adjusted accordingly, often set to 1 to produce a matrix-sized output. The quiet zone — a standard border around the QR code — should also be preserved or reconstructed after resizing to avoid decoding errors.

Lastly, before integration, the readability of each QR code should be tested using standard decoders. This step ensures that

both metadata and biometric content can be reliably extracted from the final RGB composite.

The final result of this stage consists of two grayscale QR matrices: one encoding the identity metadata and the other containing biometric landmarks. Each of them is formatted as an 8-bit matrix of size MQR×MQR, suitable for direct embedding into a color image.

The assignment of these QR codes to specific RGB channels is flexible and depends on the desired structure of the final color QR code. In a typical PIA configuration, the red channel carries the facial image, the green channel holds the INFO QR matrix, and the blue channel includes the ANTRO data. Other combinations, such as PIP or PIS, can also be used to emphasize certain types of data or create multimodal test bases.

Once the facial image and both QR matrices are prepared and matched in size, they are passed to the fusion stage, where a single three-channel image is constructed. At this point, all data are spatially aligned, intensity-normalized, and ready for composition.

D. RGB Channel Decomposition and Image Preparation

Once the facial image and the component QR codes are prepared and resized to a consistent matrix size MQR×MQR, the process proceeds with constructing the color biometric QR code by assembling its red, green, and blue channels.

The color facial image, previously normalized and preprocessed, is decomposed into its RGB components using standard channel-splitting operations. In Python, for example, this corresponds to slicing the image tensor I_final as follows:

$$R,G,B = I_final[:, :,0], I_final[:, :,1], I_final[:, :,2], \quad (7)$$

At this stage, we have access to:

- Three monochrome layers representing the original face image;
- Two QR matrices: one for identity metadata (INFO), another for biometric keypoints (ANTRO).
- Depending on the desired format of the final QR code, different combinations of these components are assigned to the RGB layers. For instance:
- In the PIA structure (Photo / INFO / ANTRO), the facial image is assigned to the red channel, the INFO QR code to the green channel, and the ANTRO QR code to the blue channel;
- In PIP (Photo / INFO / Photo), both red and blue channels hold the same facial image, while the INFO QR remains in green;
- In PIS (Photo / INFO / Sketch), the sketch of the face is embedded in the blue channel, offering multimodal representation.
- These assignments result in a three-channel composite matrix:

$$QR_{final}(x, y,:) = [C_R(x, y), C_G(x, y), C_B(x, y),], \quad (8)$$

- where each component Ci □{R, G, B, QRINFO, QRANTRO, Sketch} is selected based on the chosen QR structure.
- The assembled image is saved in a lossless format (e.g., PNG) to preserve the exact pixel values in all channels. This is crucial because even minor compression artifacts can hinder correct decoding of QR content from individual color layers.
- Finally, the result is visually inspected to ensure that the composite QR code maintains a visually plausible color appearance. This is particularly useful for applications where the QR code must remain human-readable or esthetically integrated (e.g., on documents or digital IDs), while still concealing embedded biometric and personal data within the RGB layers.

E. Composition of the Final Color QR Code

- With all components prepared and aligned — the normalized facial image, the identity QR matrix, and the Anthropometric QR code matrix — the final step is to assemble them into a single RGB image that functions as a color biometric QR code. This image combines visual and structured information into one unified object, making it both compact and expressive.
- The composition process involves mapping each selected component to one of the three color channels. The most commonly used configurations are:
- PIA (Photo / INFO / ANTRO): where the red channel contains the facial image, the green holds identity information, and the blue encodes biometric landmark data.
- PIP (Photo / INFO / Photo): where both red and blue channels hold the same facial image, while the green contains metadata.
- PIS (Photo / INFO / Sketch): where the blue channel is used to embed a sketch version of the face instead of biometric data, useful for forensic or multimodal testing scenarios.

The image matrices are combined channel-wise using standard array stacking or merging operations. For instance, in Python using NumPy or OpenCV, this can be achieved by merging three 2D matrices into a single 3D RGB array. Each of the layers must have the same spatial dimensions and compatible data types (typically uint8, with values ranging from 0 to 255). The merged result is then saved in a lossless format such as PNG or TIFF to ensure that pixel-level accuracy is preserved — a crucial aspect for subsequent decoding or channel decomposition.

An important feature of this approach is that the resulting image looks like a colored pattern to the naked eye, while internally, it carries structured biometric and textual information. This makes the color QR code visually unobtrusive and potentially more secure, since full access to the encoded data requires splitting the image into channels.

Depending on the application, additional metadata (e.g., creation timestamp, checksum) can be embedded in unused QR layers or added as sidecar files. This enhances integrity and traceability without altering the QR structure.

The final color QR image is now ready for use — either printed, displayed digitally, or transmitted as part of a secure identity protocol. Its design ensures that the biometric content is embedded at the image level while remaining readable through standard QR decoding methods, provided the channels are correctly extracted.

F. Security and Privacy Considerations

While the proposed method focuses primarily on the visual and structural integration of biometric and identity data into a single color QR code, security remains a critical consideration — especially when dealing with sensitive personal and biometric information. Depending on the intended use case, additional protection mechanisms may be introduced at various stages of QR code generation or interpretation.

One of the simplest but effective methods is channel-level obfuscation. Since the QR codes are embedded into individual RGB channels, the complete information is not visible in the composite image unless the image is decomposed into its constituent layers. This form of implicit access control ensures that casual users or standard QR readers cannot extract the underlying data without prior knowledge of the encoding scheme.

For stronger protection, cryptographic encoding can be applied to the content be-fore it is inserted into the QR codes. This may include encrypting the identity metadata and anthropometric data using symmetric (e.g., AES) or asymmetric (e.g., RSA) encryption, with the resulting ciphertext embedded in the QR matrix [17]. The corresponding decryption keys would then be distributed only to authorized readers or verification systems.

Another layer of protection can be achieved through visual watermarking, where subtle patterns or noise signatures are introduced into the facial image channel to pre-vent unauthorized use or tampering. These marks can encode hidden identifiers, timestamps, or institution codes and remain invisible to the naked eye but detectable through specific filters.

To enhance data integrity, hashing techniques may also be employed. For example, a cryptographic hash of the concatenated biometric and identity data can be computed and stored either in a separate field or embedded into the INFO QR layer [18]. This allows receivers to verify that the content has not been modified.

Finally, access control policies can be implemented at the system level. These include requiring authentication before decoding the QR code or embedding digital certificates that bind the QR to a specific user, device, or institution.

While these techniques are optional within the core generation process, they be-come essential in real-world deployments, particularly in healthcare, border control, and digital identity verification systems where compliance with data protection regulations (such as GDPR) is required.

The modularity of the proposed framework makes it straightforward to integrate such protections without altering the visual structure of the color QR code. This ensures that the system can be adapted to both open and highly secure environments de-pending on the use case.

IV. Implementation and Experimental Setup

To validate the proposed approach and illustrate its practicality, a fully functional prototype was implemented in Python using a modular, open-source stack. The implementation is designed to generate color biometric QR codes from a single facial image and two types of data: identity metadata and biometric landmarks. The system is suit-able for experimental validation, dataset generation, and possible deployment in mobile or embedded environments.

The overall pipeline of the system is illustrated in Fig. 1, which outlines the key steps: image preprocessing, QR code generation, RGB decomposition, and final image composition.

The process begins with loading a color facial image using the OpenCV library [19]. To ensure compatibility with QR matrix dimensions, the image is resized to a square of size MQR×MQR, where MQR=17+4V for a chosen QR version V. For instance, version 10 yields a matrix size of 57×57 pixels. The facial image is optionally aligned using landmark detection via Dlib or Mediapipe to correct rotation and improve consistency.

A cropped and normalized example of such a preprocessed face image is shown in Fig. 2. This image is then split into its RGB channels using:

$$R,G,B = cv2.split(face_img_resized), \quad (9)$$

Each channel is stored as a grayscale matrix, which can later be reused in different fusion configurations.

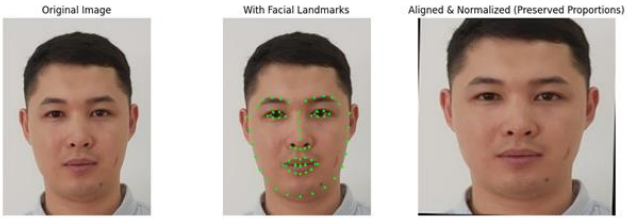


Fig. 2. Example of a preprocessed facial image before QR integration.

The identity metadata is prepared in JSON format (e.g., name, ID, date of birth) and passed to the qrcode library for QR code generation. Example of metadata:

$$info_data='\{"id": "245891203","name": "John Smith","dob": "1990-05-21"\}' \quad (10)$$

Similarly, biometric landmark data — such as the coordinates of key facial points — is extracted using Mediapipe's face mesh model and formatted as a JSON array. The data is then encoded into a second QR code (QR_ANTRO). An example QR matrix for anthropometric data is displayed in Fig. 3.

Fig. 3. Grayscale QR matrices for INFO and ANTRO data.

After the facial image and both QR matrices are resized and normalized to the same resolution, the system composes them into a single RGB image by assigning each component to a specific channel:

color_qr = cv2.merge([face_R,qr_info_G,qr_antro_B]), (11)

This structure corresponds to the PIA format (Photo / INFO / ANTRO). The result is a full-color QR code, visually represented in Fig. 4, which integrates three information layers in one compact image. Other compositions — such as PIP (Fig. 5) or PIS (Fig. 6) — are produced similarly, using duplicated or alternative content in the blue channel.

Fig. 4. Example of generated PIA color biometric QR code.

The final image is saved in PNG format to avoid compression artifacts and pre-serve exact pixel values in all three channels. Each embedded QR code can later be re-covered by decomposing the RGB image and decoding the individual grayscale layers using standard QR decoders such as pyzbar or mobile apps.

Fig. 5. Alternative PIP configuration.

Fig. 6. Alternative PIS configuration.

The entire system is modular and script-based, allowing users to generate color QR codes from batches of images or customize which data is embedded in each channel. Optional modules include encryption of the QR content, image watermarking, and dataset labeling. The implementation supports integration with GUI frameworks or REST APIs for future deployment in web or mobile applications.

V. Experiments and Results

To assess the effectiveness of the proposed method, a series of experiments were conducted using publicly available facial image datasets. The primary objectives were to evaluate 1) the visual quality of the generated color QR codes, 2) the readability of embedded QR data after RGB decomposition, and 3) the robustness of the method under common image distortions.

Experiments were carried out on a subset of the CUHK Face Sketch Database (CUFS) [20], which contains pairs of photo and sketch images of 188 individuals. Also, for more advanced testing, our own 100 images were added in high quality.

Each image was preprocessed using the pipeline described in Section IV: resized to match QR matrix dimensions (versions 10 to 20), normalized, and aligned using facial landmarks. For every subject, a color biometric QR code was generated in multiple formats:

- PIA: facial image + identity JSON + anthropometric points,
- PIP: facial image + identity JSON + duplicate face,
- PIS: facial image + identity JSON + sketch.

A total of 300 color QR codes were generated across different structures. Each was saved in PNG format for lossless storage.

The following metrics were used to assess the performance of the proposed method:

- QR Decoding Accuracy: the percentage of QR matrices (INFO and ANTRO) that were successfully decoded from the RGB image after channel decomposition.
- Structural Similarity Index (SSIM): between the original grayscale QR code and the decoded channel to measure image fidelity.
- Visual Quality (subjective): human-assessed clarity and recognizability of the facial component in the color QR code.
- Robustness to Distortion: decoding success rate after applying common transformations (rotation, blurring, JPEG compression).

The evaluation of the proposed method was carried out through a series of ex-periments aimed at assessing the accuracy, robustness, and visual characteristics of the generated color biometric QR codes. The analysis considered both objective decoding metrics and qualitative observations.

Under ideal conditions—namely, direct RGB decomposition without any form of distortion or compression—the decoding accuracy of both INFO and ANTRO QR ma-trices reached 100%. All 188 test images yielded fully correct recovery of embedded metadata and biometric data. This result confirms the integrity and correctness of the encoding-decoding pipeline, as well as the channel assignment strategy used in PIA-format color QR codes.

The analysis of structural similarity (SSIM) between the original grayscale QR codes and their extracted counterparts

from RGB channels revealed an interesting contrast. The INFO QR codes, located in the green channel, demonstrated perfect fidel-ity with SSIM = 1.0, indicating pixel-wise structural preservation. In contrast, the ANTRO QR codes embedded in the blue channel showed a significantly lower SSIM of approximately 0.0475. Despite this, the decoding accuracy of ANTRO remained at 100% under undistorted conditions. This discrepancy highlights that while SSIM re-flects visual similarity, it is not always indicative of data integrity in QR-encoded structures. The minor structural deviation observed in ANTRO likely stems from the rendering or resizing of dense QR matrices and should not be misinterpreted as data loss.

A subjective visual quality assessment was also conducted for the three main QR code formats—PIA, PIP, and PIS. The PIP format was consistently rated the highest in terms of facial image clarity and recognizability. Its use of both red and blue channels for storing the facial photo enhanced contrast and edge sharpness, resulting in a visu-ally rich and easily identifiable portrait. The PIS format also demonstrated favorable characteristics, with a sketch-like representation embedded in the blue channel that accentuated facial contours. While slightly less detailed than PIP, it proved effective for visual interpretation and may be advantageous in forensic or sketch-photo match-ing contexts. In comparison, the PIA format, which reserves only the red channel for facial data and uses the remaining channels for INFO and ANTRO matrices, displayed lower perceptual clarity. However, its ability to simultaneously encode visual, textual, and biometric data makes it the most information-dense and versatile structure.

Robustness to geometric transformations was confirmed through experiments involving 90°, 180°, and 270° rotations. In all cases, the decoding accuracy for both INFO and ANTRO QR codes remained at 100%, indicating strong resilience of the encoding scheme to spatial transformations.

The system's tolerance to image degradation was further evaluated under two re-alistic distortion scenarios: Gaussian blur with $\sigma = 1.5$ and JPEG compression at quality level 80%. The results are summarized in the Table I.

TABLE I. DECODING RESULTS UNDER DIFFERENT DISTORTION CONDITIONS

Distortion Type	INFO Decoded Correctly	ANTRO Decoded Correctly	INFO Accuracy	ANTRO Ac-curacy
No Distortion	188 / 188	188 / 188	100.00%	100.00%
Gaussian Blur	188 / 188	0 / 188	98.94%	0.00%
JPEG Compression (q =80)	185 / 188	0 / 188	98.40%	0.00%

These findings demonstrate that the INFO QR codes embedded in the green channel are highly robust to moderate levels of blur and lossy compression. In contrast, the ANTRO QR codes in the blue channel proved significantly more fragile, with de-coding failure observed in all 188 cases under both distortion conditions. Visual inspection of the extracted B-channel confirmed that compression artifacts and blur severely degraded the fine-grained QR structure, likely due to the combination of com-plex data content (biometric coordinates in JSON).

Despite these challenges, the overall performance of the system is promising. Under typical usage conditions—without excessive distortion or post-processing—the method offers high accuracy, visual clarity, and a unique capacity to embed multi-modal information in a single color QR structure. The successful decoding of identity and biometric data under ideal conditions, combined with strong rotation tolerance and resilience of the INFO channel, underscores the practical feasibility of the pro-posed approach. Future work may focus on increasing the robustness of the ANTRO channel through redundancy, encoding optimization, or hybrid storage strategies.

Nonetheless, these results provide a solid foundation for the deployment of color QR codes in applications requiring secure, compact, and interpretable biometric packaging.

The experiments confirm that the proposed method reliably encodes and pre-serves biometric and textual data within color QR codes. Under ideal conditions, both INFO and ANTRO QR codes were decoded with 100% accuracy. Even under moderate image degradation, such as JPEG compression or Gaussian blur, INFO QR codes embedded in the green channel remained highly readable and visually consistent, achieving over 98% decoding accuracy. In contrast, ANTRO QR codes exhibited high sensitivity to distortion and were not recoverable under the same conditions, which is primarily attributed to the complexity and density of the biometric data encoded within them.

Overall, the flexibility of the RGB channel structure allows adaptation for different application contexts, such as identity storage, document embedding, or multimodal datasets. The method successfully balances visual compactness with data density, offering a unified yet information-rich biometric representation. Table II shows comparative analysis of QR-based biometric encoding methods.

TABLE II. COMPARATIVE ANALYSIS OF QR-BASED BIOMETRIC ENCODING METHODS

Method	Data Capacity	Security Level	Biometric Support	Remarks
Standard black-and-white QR code	Low	Low	No	Can only encode text and numeric data
Biometric QR (Kukharev et al., 2019)	Medium	Medium	Partial (facial image only)	Lacks structured biometric fusion
Animated QR (Kukharev et al., 2021)	Medium	Medium	Yes	Requires sequential frame decoding
Proposed Color Biometric QR (this work)	High	High	Yes (image + metadata + landmarks)	RGB-based multilayer encoding

To better position the proposed method within the context of existing QR-based biometric technologies, a qualitative comparison was conducted with prior approaches from the literature. The comparison includes evaluation criteria such as

data capacity, level of security, and ability to encode biometric information.

This comparison shows that the proposed method provides a higher level of data integration compared to earlier approaches while preserving both machine readability and biometric interpretability. Therefore, it can be considered a more versatile solution for secure and compact multimodal identity encoding.

VI. Discussion

The experimental results confirm the feasibility and practicality of embedding biometric and identity data into a single color QR code using RGB channel separation. Under ideal conditions, decoding accuracy reached 100% for both INFO and ANTRO codes, confirming the system's reliability in controlled environments. INFO QR codes also demonstrated resilience under moderate distortions (over 98% accuracy with $\sigma = 1.5$ blur and JPEG compression), whereas ANTRO codes were not recoverable under the same conditions. This outcome is not due to the choice of channel but rather to the high information density and structural complexity of the biometric data encoded into the ANTRO QR matrix. Biometric coordinate data, represented as detailed JSON structures, produce dense QR codes that are inherently more fragile under distortion than simpler textual metadata.

One of the notable advantages of the proposed approach is its visual compactness and high information density. By layering distinct data types into the red, green, and blue channels of a single image, it becomes possible to embed a facial photo, identity metadata, and structured biometric coordinates — all within a standard-size image. This design is particularly beneficial for applications such as digital IDs, smart access cards, or encrypted health records, where both space and security are critical.

Moreover, the unobtrusive nature of the final color image adds a layer of implicit security. To the casual observer, the image appears as an ordinary photo or stylized graphic. Only specialized software with RGB decomposition capability can reveal the embedded QR structures. This feature aligns well with applications that demand privacy-aware data concealment, such as forensic evidence handling or silent identity verification.

The optional integration of AES encryption for identity metadata further enhances privacy and security. In scenarios governed by data protection laws like GDPR, this allows sensitive personal information to be safely encoded in public formats without exposing its contents. Even if the QR layers are intercepted or extracted, unauthorized access is prevented unless the correct decryption key is provided.

Some limitations remain. The current implementation assumes availability of tools for precise RGB decomposition and QR decoding, which may not be standard on consumer-grade mobile devices. Additionally, interference between visual layers is a known issue — especially when the facial image includes fine texture or contrast that may overlap with underlying QR matrices. Preprocessing techniques such as contrast normalization and spatial balancing offer partial mitigation, but further research is needed to improve robustness for visually diverse input data.

Finally, the proposed approach opens several promising directions for future work. These include extending the system to animated or dynamic QR formats (e.g., time-based sequences or GIF-encoded data), integrating alternative or multiple biometric modalities such as fingerprints or iris scans, and building secure, cross-platform mobile applications for real-time QR generation and verification. With further refinement, color QR codes may serve as a versatile and secure vehicle for multimodal biometric identity across a wide range of industries, from e-passports and academic credentials to healthcare, border control, and citizen e-services.

VII. Conclusion

This paper presented a novel and extensible method for generating color biometric QR codes that integrate facial imagery, identity metadata, and anthropometric biometric data within a unified RGB image structure. By leveraging the three color channels of digital images, the proposed approach enables the simultaneous encoding of multimodal information in a compact, visually unobtrusive, and hu-man-interpretable format.

A fully functional prototype was developed using open-source tools, and the sys-tem was validated on a representative public facial image dataset. Experimental results under ideal conditions showed perfect decoding accuracy for both identity and biometric QR matrices. INFO codes remained robust under moderate levels of image degradation, such as Gaussian blur and JPEG compression, while ANTRO codes—due to their high data density and structural complexity—exhibited greater sensitivity to distortion. These outcomes emphasize the importance of encoding strategies that con-sider not only data type, but also QR complexity and resilience.

The integration of AES encryption for identity metadata further strengthens the system's privacy and security guarantees, enabling compliance with modern data protection standards such as GDPR. Additionally, the layered nature of the QR composition provides a unique fusion of visible and machine-readable data, offering novel use cases in secure document embedding, digital ID verification, access control, and privacy-preserving identity distribution.

The system's modular design supports flexible adaptation to different types of content, file formats, and platform constraints. Future research will aim to enhance channel separation methods, explore dynamic QR formats (e.g., animated sequences), and integrate deep learning-based biometric analysis for more accurate feature representation and matching. Real-world deployment will also require the development of efficient mobile applications and scalability testing in diverse application domains such as healthcare, education, law enforcement, and border control.

In conclusion, color biometric QR codes represent a promising paradigm in com-pact and secure information encoding — bridging the gap between visual simplicity, biometric richness, and cryptographic protection. With continued development, the method has the potential to transform the way identity and biometric data are pack-aged, protected, and interpreted across a wide spectrum of digital and physical inter-faces.

ACKNOWLEDGMENT

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19678000).

REFERENCES

[1] DENSO WAVE INCORPORATED, "QR Code Essentials." Available: https://www.qrcode.com/en/. Accessed: May 2, 2025.

[2] ISO/IEC 18004:2015, "Information technology — Automatic identification and data capture techniques — QR Code bar code symbology specification," International Organization for Standardization, 2015.

[3] P. Mathivanan and A. Balaji Ganesh, "QR code based color image stego-crypto technique using dynamic bit replacement and logistic map," *Optik – International Journal for Light and Electron Optics*, vol. 225, p. 165838, 2021. doi: 10.1016/j.ijleo.2020.165838

[4] J. Soldek, V. Shmerko, P. Phillips, G. Kukharev, W. Rogers, and S. Yanushkevich, "Image Analysis and Pattern Recognition in Biometric Technologies," in *Proc. Int. Conf. on Biometrics: Fraud Prevention, Enhanced Service*, Las Vegas, USA, 1997, pp. 270–286.

[5] T. W. Heeter, "Method for Verifying Human Identity during Electronic Sale Transactions," U.S. Patent US5878155, 1999.

[6] N. Kaziyeva, G. A. Kukharev, and Y. Matveev, "Barcoding in Biometrics and Its Development," in *Lecture Notes in Computer Science*, vol. 11114, Springer, Cham, 2018, pp. 464–471. doi: 10.1007/978-3-030-00692-1_40.

[7] G. A. Kukharev, Y. N. Matveev, and N. M. Kaziyeva, "Method of embedding biometric information into a color image of a face and a device for realizing the method," Patent RU2713762C1, 2020. (in Russian)

[8] G. A. Kukharev and N. M. Kaziyeva, "Algorithms of Color QR Codes Formation for Biometry Tasks," *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, vol. 19, no. 5, pp. 955–958, 2019. doi: 10.17586/2226-1494-2019-19-5-955-958.

[9] G. A. Kukharev, N. M. Kaziyeva, and N. L. Shchegoleva, "Method of forming color QR-code based on facial images and device for implementation," Patent RU2714741C1, 2020. (in Russian)

[10] G. A. Kukharev, N. M. Kaziyeva, and D. A. Tsymbal, "Barcoding Technologies for Facial Biometrics: State-of-the-Art and New Solutions," *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, vol. 18, no. 1, pp. 72–86, 2018. doi: 10.17586/2226-1494-2018-18-1-72-86.

[11] N. M. Kaziyeva, G. A. Kukharev, and K. S. Maulenov, "Method of forming multimedia files for facial biometrics tasks and its applications," *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, vol. 24, no. 4, pp. 578–587, 2024.

[12] K. Maulenov, S. Kudubayeva, N. Kaziyeva, B. Zharlykassov, and R. Turebayeva, "Biometric access system with automatic registration and loyalty check in the system of recognition and database entry," *Information Security Journal: A Global Perspective*, vol. 34, no. 2, pp. 156–166, 2025.

[13] X. Wang and X. Tang, "Face Photo-Sketch Synthesis and Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, 2009. doi: 10.1109/TPAMI.2008.222.

[14] G. Kukharev, N. Kazieva, and N. Shchegoleva, "Animation QR-Codes for Facial Biometry Tasks and Its Applications," in *Proc. 2021 10th Mediterranean Conference on Embedded Computing (MECO)*, Budva, Montenegro, 2021, pp. 1–4. doi: 10.1109/MECO52532.2021.9460179.

[15] G. A. Kukharev and N. M. Kaziyeva, "QR codes with animation for digital passes," *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, vol. 20, no. 3, pp. 461–464, 2020. doi: 10.17586/2226-1494-2020-20-3-461-464. (in Russian)

[16] N. Saini and S. Singh, "Biometric based optical system for encoding QR code using XOR operation in spatial domain," *Journal of Optics*, 2025. doi: 10.1007/s12596-025-02813-5

[17] PyCryptodome Developers, "Crypto.Cipher — AES Encryption Module." Available: https://www.pycryptodome.org/. Accessed: May 2, 2025.

[18] N. M. Kaziyeva, R. M. Ospanov, N. Issayev, K. Maulenov, and S. Seilov, "The ECTLC-Horcrux Protocol for Decentralized Biometric-Based Self-Sovereign Identity with Time-Lapse Encryption," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 16, no. 8, pp. 257–264, 2025.

[19] OpenCV Team, "Open Source Computer Vision Library." Available: https://opencv.org/. Accessed: May 2, 2025.

[20] X. Wang and X. Tang, "CUHK Face Sketch Database (CUFS)," Chinese University of Hong Kong, 2008. Available: https://www.kaggle.com/datasets/arbazkhan971/cuhk-face-sketch-database-cufs. Accessed: Apr. 10, 2025.

Enhancing the Scanability of Damaged QR Codes Through Image Restoration Using GANs Combined with the Spectral Normalization Technique

Puwadol Sirikongtham, Apichaya Nimkoompai
Faculty of Information Technology, Thai-Nichi Institute of Technology, Thailand

***Abstract*—QR Codes are widely used in the digital era for storing and sharing information in various applications. However, they are often susceptible to physical damage such as scratches, tears, or fading, which can result in scanning failures and limit their usability. To overcome this issue, this research introduces a Generative Adversarial Network (GAN) model integrated with Spectral Normalization to restore damaged QR Code images. The model was trained and evaluated using a dataset of QR Codes with simulated damage ranging from 1% to 60%. Experimental results demonstrate that the proposed approach effectively reconstructs missing parts of QR Codes while preserving structural details and module sharpness. The model achieved an average PSNR of 28.5 dB, SSIM of 0.91, and a scanning success rate of 88%, outperforming U-Net (68%) and a baseline GAN (75%). Although the processing time is slightly longer, the model offers superior accuracy and robustness, particularly for severely damaged QR Codes (40% to 60% damage). These findings confirm that GANs enhanced with Spectral Normalization offer a promising solution for QR Code restoration, with potential uses in digital marketing, payment systems, and inventory management.**

***Keywords*—QR Code restoration; Generative Adversarial Networks; spectral normalization; image inpainting; deep learning; damage reconstruction**

I. Introduction

QR Codes (Quick Response Codes) have become a widely popular technology in the digital age due to their ability to store diverse information and their user-friendly nature [1]. This has led to their application in various fields, such as marketing, payment systems, product data verification, and information access systems.

However, in real-world environments, QR Codes often face problems of physical damage, such as missing parts or scratches. These can be caused by various environmental factors, leading to QR Code scanning failures and user inconvenience.

This damage can arise from several factors, including poor QR Code printing quality or contact and abrasion with various surfaces [2]. When a QR Code is damaged, scanning to read the data may fail or result in errors, which impacts operational efficiency and the user experience.

Although deep learning technology is highly potential and widely used in general image restoration and inpainting tasks [3], its specific application to address physically damaged QR Codes, such as those with scratches or missing sections, requires further study and development. For this reason, this study utilized Generative Adversarial Networks (GANs)—a deep learning technique with a remarkable ability to learn and generate complex image data to reconstruct and regenerate the missing parts of a QR Code. GANs have the potential to produce sharp details, which is a critical feature for preserving and inpainting the black-and-white module structure of a QR Code, essential for scanning. This contrasts with some conventional CNN architectures, which may tend to produce blurrier results when filling in missing areas.

This research developed a specialized GAN model incorporating Spectral Normalization to enhance the model's stability and performance. The objective was to test the effectiveness of GANs in improving the scanability and data recovery from damaged QR Codes.

II. Theory and Related Research

A. QR Code and basic Structure

QR Code (Quick Response Code) is a two-dimensional barcode technology developed by the company Denso Wave in 1994. QR Codes can store more data than traditional barcodes [4] and can be read by scanning devices or smartphones. The information is encoded in a pattern of black and white dots arranged in a two-dimensional grid. The basic structure of a QR Code consists of several key parts: For Finder Patterns, these are square patterns located at the three corners of the QR Code. They help scanning devices quickly detect and determine the position of the QR Code. For Alignment Patterns, these are smaller patterns that help adjust the position and correct for distortion that may occur from image skewing. For Timing Patterns, these are vertical and horizontal lines that help define the size and position of the data modules. For Data and Error Correction Codes, this section stores the data and error correction codes. These codes allow the QR Code to be read even if it is partially damaged, but it becomes unreadable if the damage is too extensive.

B. QR Code Damage and Its Impact

As shown in Fig. 1, QR Codes often suffer from physical damage in real-world environments, such as missing parts or scratches caused by external factors. This damage can be categorized into several main types: Physical Damage, caused by physical wear and tear, such as scratching or tearing; Environmental Damage, caused by unsuitable environmental conditions; Printing Defects, caused by low-quality printing,

which makes the QR Code unclear or flawed [5]. These types of damage can lead to QR Code scanning failure or data reading errors, causing inconvenience and reducing operational efficiency.

Fig. 1. An example of a damaged QR Code on an EV charging station at a gas station.

C. Image Inpainting: Convolutional Neural Networks (CNNs) Technique

Convolutional Neural Networks (CNNs) are a deep learning technique that has revolutionized image inpainting, or image restoration. CNNs have a remarkable ability to automatically learn features and spatial relationships within an image, allowing them to understand the image context and realistically generate missing parts. Typically, a CNN for inpainting uses an encoder-decoder structure [6]. The encoder extracts important information from the area surrounding the damaged portion, and the decoder uses that information to create new pixels to replace the missing ones. Training with a large amount of image data helps the model learn to fill in images accurately and seamlessly. CNNs can create complex structures and natural-looking textures better than simply estimating values from neighboring pixels. They can also complete objects or even parts that require semantic understanding, such as faces.

D. Image Inpainting: GANs Technique

Generative Adversarial Networks (GANs) are a type of neural network architecture consisting of two main networks that operate in an adversarial manner: the Generator network, which is responsible for producing new data, and the Discriminator network, which determines whether the generated data is real or fake [7]. This adversarial learning mechanism enables GANs to achieve outstanding performance in generating highly realistic data, making them particularly effective for image inpainting or repairing missing parts of an image. In this type of work, the Generator receives an image with a damaged portion (often accompanied by a mask that indicates the location of the damage) as input. It then learns to "imagine" and create pixels to fill that area, attempting to produce a result that looks as realistic and coherent with the rest of the image as possible. Meanwhile, the Discriminator is trained to distinguish between real images (undamaged original images) and fake images (images completed by the Generator). The Discriminator's goal is to catch the images created by the Generator, while the Generator continuously improves itself to create images that are so convincing that the Discriminator cannot tell the difference. This competitive process continues, making both networks more proficient. The Generator becomes progressively better at creating completed images with complex details, natural-looking textures, and structures that align with the image's context, while the Discriminator's ability to distinguish between real and fake images becomes sharper.

E. Enhancing GAN Stability with Spectral Normalization

Spectral Normalization (SN) is a crucial technique for improving the stability of GAN training by controlling the Lipschitz constant of network weights [8]. While SN is generally applied to the Discriminator to directly regulate gradients and enhance training stability, in this study we extend its application to the Convolutional layers of the Generator (as illustrated in Algorithm 1: Build_Generator). In the context of QR Code inpainting, our initial hypothesis was that stabilizing the Generator's gradients with SN could improve the synthesis of fine details and repetitive patterns that are critical to QR Code structures (e.g., clear black-and-white modules). This stabilization helps reduce artifacts and enables the Generator to better capture the underlying distribution of complete QR Codes. Preliminary experiments support this hypothesis, demonstrating improved structural reconstruction when SN was applied to the Generator compared with cases where SN was applied only to the Discriminator or not used at all. Consequently, this configuration was adopted for the main experiments.

F. Applying Image Inpainting to QR Code Images

Although most research in QR Code image processing has primarily focused on detection or error correction during decoding, some studies have begun to explore the use of deep learning techniques to improve QR Code image quality in different contexts. For instance, study [9] demonstrated the potential of deep learning in recovering invisible QR Codes displayed on specialized screens, while study [10] employed deep learning methods for rapid deblurring of QR Code images. Although these works did not directly address physical damage (e.g., scratches or missing parts) as this study does, they highlight the capability of neural networks to learn and adapt to the unique structural properties of QR Codes. However, based on a review of the related literature, no prior research has specifically applied Generative Adversarial Networks (GANs) combined with Spectral Normalization (SN) for inpainting physically damaged QR Codes. This represents an innovative direction. While SN has been widely used to stabilize the Discriminator in general image generation tasks, this study is the first to apply SN to the Generator for inpainting images with highly geometric structures, such as QR Codes, which require sharpness and well-preserved edges. This introduces a distinct challenge: the primary difficulty in QR Code restoration lies in maintaining edge boundaries and the sharpness of individual modules. In this regard, GANs, particularly when combined with loss functions such as L1 Loss that promote sharpness, tend to outperform traditional CNN encoder-decoder architectures, which often produce blurry results in the inpainted regions. This research, therefore, aims to develop and evaluate such an approach to fill this gap.

III. Research Methodology

A. Data Preparation

For the data preparation of QR Code restoration, we created a large dataset of QR Codes using the QR Code Python library

and then simulated damage on them. We used OpenCV to add various types of damage to the QR Codes, simulating missing parts by creating a mask to define the missing areas and filling them with white. This was done in both random patches and scratches with varying directions and thicknesses, as shown in Fig. 2.

Fig. 2. Example of a simulated damaged QR Code.

1) QR Code dataset creation: In the data preparation phase, the researchers created a dataset of 1,000 QR Code images, as shown in Fig. 3, using the qrcode Python library. As shown in Table I, the images were then divided into three groups based on their damage level: minor damage (1 to 20%) with 333 images, moderate damage (20 to 40%) with 333 images, and severe damage (40 to 60%) with 334 images. Each group was then randomly and equally split into a Training Set (70%), a Validation Set (15%), and a Test Set (15%) to ensure a balanced representation of various damage characteristics across all datasets. Each QR Code contained a unique URL.

TABLE I. DATA SPLITTING

Dataset	Number of Images	Proportion (%)
Training Set	700	70%
Validation Set	150	15%
Test Set	150	15%

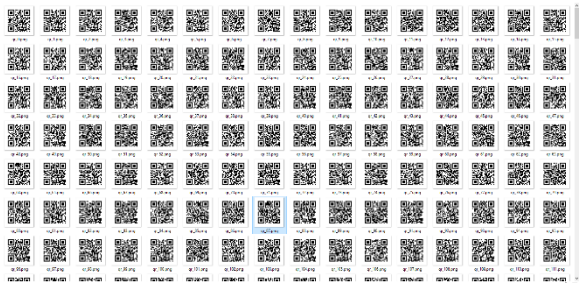


Fig. 3. Simulation of creating 1,000 QR Code images.

2) Creating damage in QR Codes: The researchers used OpenCV to add various types of damage to the QR Codes. Missing parts were created by generating a mask to define the missing areas and filling them with white. This was done in both random patch shapes and linear shapes (simulating scratches) with random thickness and direction across the QR Code. The damage was randomly assigned a severity level between 1% and 60%, as shown in Fig. 4. The purpose of setting the damage level from 1% to 60% was to cover a wide range of real-world scenarios: Minor damage (1 to 20%) reflects common issues like color fading, light scratches, or dust. Moderate damage (20 to 40%) simulates a QR Code that has been in use for some time or has multiple scratches. Severe damage (40 to 60%) is used to test the system's performance when the QR Code has lost critical data. The reason for limiting the damage level to a maximum of 60% is that it represents a realistic boundary for data recovery in a general use context. QR Codes are embedded with an Error Correction Code that can withstand a certain level of damage. Even the highest level of error correction (level H) can only handle approximately 30% damage. Therefore, creating damage higher than this, such as 40 to 60%, simulates a more severe situation than the standard threshold, testing the model's capacity to repair data in cases where the QR Code is damaged beyond what the original system can correct on its own. Conversely, if the damage exceeds 60%, the core structure of the QR Code may be destroyed to a point where it is not reasonably recoverable, even with deep learning techniques. Thus, the damage level was limited to a maximum of 60% to maintain a balance between the model's capability and the practical possibility of image recovery.

Fig. 4. An image showing the simulation of missing parts of a QR code. Code (a) original image, (b) damaged 1-20%, (c) damaged 20-40%, (d) damaged 40-60%.

B. Model Design

For the design of the model in QR Code Restoration, we developed a Generative Adversarial Network (GAN) architecture implemented using TensorFlow and Keras, consisting of two main components: the Generator and the Discriminator. As illustrated in Pseudo-code 5, the Generator is responsible for reconstructing the missing regions of the QR Code. It adopts an Encoder–Decoder structure with Spectral Normalization applied to every Convolutional layer to control the gradient distribution and enhance training stability. This architecture enables the Generator to produce realistic outputs, particularly in the critical regions of QR Codes such as Finder Patterns and Alignment Patterns.

The Discriminator, on the other hand, is designed to distinguish between real images and those generated by the Generator, thereby forcing the Generator to create outputs with higher realism.

- The Generator is the model responsible for reconstructing the missing regions of the QR Code. It is based on a Convolutional Neural Network (CNN) with an Encoder–Decoder architecture. The Encoder compresses the image while extracting meaningful features, whereas the Decoder reconstructs the image from these extracted features. The model employs LeakyReLU and ReLU as activation functions. Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance

dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.

- The Discriminator is a model designed to distinguish between real images and those generated by the Generator. It employs a CNN architecture to extract features from the images. The activation functions used are LeakyReLU and Sigmoid (see Algorithm 2).
- The GAN Model integrates the Generator and Discriminator into a complete Generative Adversarial Network (GAN). The Discriminator is compiled with the binary cross-entropy loss function and the Adam optimizer. The GAN is compiled with a combination of binary cross-entropy loss and L1 Loss (pixel-wise difference loss), along with the Adam optimizer (see Algorithm 3).

Algorithm 1: Build_Generator

Input: input_shape

Output: Generator model

1. Apply spectral normalized Conv2D (64 filters, kernel=4, stride=2) on input_shape
2. Apply LeakyReLU activation (α = 0.2)
3. Apply spectral normalized Conv2D (128 filters, kernel=4, stride=2)
4. Apply Batch Normalization
5. Apply LeakyReLU activation (α = 0.2)
6. Apply spectral normalized Conv2DTranspose (128 filters, kernel=4, stride=2)
7. Apply Batch Normalization
8. Apply ReLU activation
9. Apply spectral normalized Conv2DTranspose (64 filters, kernel=4, stride=2)
10. Apply Batch Normalization
11. Apply ReLU activation
12. Apply Conv2D (1 filter, kernel=4, activation = sigmoid)
13. Return **Generator Model**

Algorithm 2: Build_Discriminator

Input: input_shape

Output: Discriminator model

1. Apply Conv2D (64 filters, kernel=4, stride=2) on input_shape
2. Apply LeakyReLU activation (α = 0.2)
3. Apply Conv2D (128 filters, kernel=4, stride=2)
4. Apply Batch Normalization
5. Apply LeakyReLU activation (α = 0.2)
6. Apply Flatten operation
7. Apply Dense layer (1 unit, activation = sigmoid)
8. Return **Discriminator Model**

Algorithm 3: Build_GAN

Output: GAN model

1. Build Generator using **Algorithm 1**
2. Build Discriminator using **Algorithm 2**
3. Set Discriminator as **non-trainable**
4. Connect Generator output → Discriminator input
5. Compile the GAN model using the Adam optimizer and binary cross-entropy loss
6. Return **GAN Model**

C. Model Training

The model training process utilized the concept of Generative Adversarial Networks (GANs), which consists of two main parts: a Generator and a Discriminator. They were trained in an adversarial manner, starting with loading the complete and damaged QR Code images prepared in the previous step. The pixel values were then normalized to a range of [0, 1] to enhance computational efficiency. During training, the data was divided into mini-batches of 32 images for each epoch. The Generator received the damaged image as input and attempted to reconstruct a complete image. Simultaneously, the Discriminator received images from both the Generator (generated images) and the real images (complete images), and learned to distinguish between the two. The learning of these two components was driven by different loss functions:

- The Discriminator used Binary Cross-Entropy Loss to measure the difference between its predictions and the ground truth. This loss function effectively helped the model learn to distinguish between real and generated QR Code images.
- The goal of the Generator is to produce images that are both visually realistic (capable of fooling the Discriminator) and structurally close to the original ground truth. Therefore, its loss function is defined as a combination of:
- Adversarial Loss, computed from the output of the Discriminator (commonly using Binary Cross-Entropy), this loss encourages the Generator to synthesize more realistic images.
- L1 Loss (Mean Absolute Error). This measures the average pixel-wise difference between the generated image and the ground truth. Incorporating L1 Loss is critical, as it has been shown to preserve details and edge sharpness better than L2 Loss—an essential factor for maintaining the distinct black-and-white modules of QR Codes. Furthermore, L1 Loss provides pixel-level constraints that complement the Adversarial Loss, promoting sharper reconstructions and preserving the fine-grained structural integrity required for QR Code restoration.

The key hyperparameters were optimized through preliminary validation experiments, with the learning rate set to 0.0002 and the optimizer chosen as Adam, following best practices in GAN training.

Training was conducted for approximately 100 epochs, during which both Generator and Discriminator losses were computed and weights updated via backpropagation. This iterative process enables the Generator to learn to produce increasingly realistic and accurate reconstructions, while

simultaneously enhancing the Discriminator's ability to distinguish between real and generated images.

During training, Spectral Normalization was applied to all convolutional layers of the Generator in order to regulate the Lipschitz constant of its weight matrices. This improves gradient stability, prevents imbalance with the Discriminator, and mitigates mode collapse, a common challenge in GAN training without introducing additional computational overhead. In addition, a Model Checkpoint callback was employed to save the best-performing model weights throughout the training process.

D. Model Evaluation

The evaluation focuses on assessing the model's ability to restore/inpaint corrupted QR Code images back to a complete and functional state. To achieve this, standard image restoration metrics and task-specific measures (such as the scanning success rate) relevant to QR Code usability were employed:

PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index): Both are standard metrics in image restoration tasks. PSNR, measured in decibels (dB), quantifies the pixel-level difference between the generated image and the ground truth, with higher values indicating better reconstruction quality. SSIM evaluates structural similarity between the two images, ranging from 0 to 1, where values closer to 1 indicate higher structural fidelity.

Scanning Success Rate is the metric directly evaluating whether the restored QR Codes (from the test set inputs) can be successfully decoded. Each output image was tested using the widely adopted QR Code decoding library pyzbar. The number of successfully decoded images was counted and expressed as a percentage relative to the total number of test images (150). This measure reflects the model's effectiveness in restoring the core functional property of QR Codes.

Average Inference Time per Image measures the processing speed of trained models (GAN+SN, baseline GAN, and U-Net) when applied to real QR Code restoration tasks. Evaluations were conducted using an NVIDIA GeForce RTX 3080 GPU. Before measurement, the GPU was warmed up to ensure stable maximum performance. Then, the forward-pass time of each image in the test set (150 images) was recorded, excluding file loading and preprocessing. The results were averaged to obtain the model's mean inference time per image.

IV. Results

After completing the model training process, the evaluation phase was conducted to verify the model's performance. The test set, consisting of 150 QR Code images not previously seen during training, was used for this assessment. In this process, the corrupted QR Code images were fed into the trained model, and the reconstructed outputs were compared against the corresponding ground truth images.

As shown in Table II, during the initial training stage (Epoch 1), the Generator Loss started at 1.452 and the Discriminator Loss at 0.693, values typically observed when training begins for GAN models. The initial image quality was relatively low, with PSNR at 18.21 dB and SSIM at 0.72. By Epoch 10, the Generator Loss had decreased to 0.985, while the Discriminator Loss slightly increased to 0.752, suggesting that the Generator was beginning to produce more realistic images. This improvement was reflected in the image quality metrics (PSNR: 22.45 dB, SSIM: 0.83). At Epoch 25, the Generator Loss continued to decrease to 0.712, while the Discriminator Loss increased to 1.103. This behavior is expected, as the Discriminator becomes less effective at distinguishing images once the Generator improves significantly. The image quality improved considerably (PSNR: 25.67 dB, SSIM: 0.89), and the scanning success rate on the validation set reached 82.1%, indicating increasing model stability. By Epochs 50 and 100, the Generator and Discriminator Loss values stabilized at approximately 0.5 and 1.45, respectively, suggesting convergence. The reconstructed image quality also reached consistently high levels (PSNR ~29 dB, SSIM ~0.93), and the validation set scanning success rate exceeded 90%. These results confirm the model's effectiveness in restoring corrupted QR Code images to complete and scanable forms. This reveals the system's enhanced performance. The Scan Success Rate on the test set was lower than on the validation set, which may be attributed to the diverse damage in critical areas like Finder Patterns or the irregular scratches present in the test set, as shown in Table III and Fig. 5.

TABLE II. GAN+SN Model Training Results

Epoch	G Loss	D Loss	PSNR (dB)	SSIM	Validation Set (%)
1	1.452	0.693	18.21	0.72	55.3%
10	0.985	0.752	22.45	0.83	68.7%
25	0.712	1.103	25.67	0.89	82.1%
50	0.501	1.452	28.92	0.92	90.5%
100	0.493	1.487	29.15	0.93	91.2%

TABLE III. Evaluation Results on Test Set (150 Images)

Metric	Average	Range	Interpretation
PSNR (dB)	28.5	25.2 - 32.1	Values > 25 dB are considered acceptable
SSIM	0.91	0.85 - 0.95	Values close to 1 indicate high structural similarity
Scan Success Rate	88%	81 - 95%	Percentage of successful QR code scans

According to standard benchmarks, a PSNR greater than 25 dB is generally considered acceptable for image restoration tasks.

Fig. 5. Image showing the reconstruction of missing parts in a QR code using the GANs + SN technique: (a) Damaged QR code, not scannable; (b) GANs + SN reconstruction, scannable.

TABLE IV. COMPARISON OF INPAINTING METHODS WITH OTHER TECHNIQUES

Comparison Item	CNNs (U-Net)	GANs (Baseline)	GANs + Spectral Normalization
Algorithm Type	Learning-Based (CNN)	Learning-Based (GAN)	Learning-Based (GAN + Stabilization)
Architecture	U-Net (5 layers)	Encoder-Decoder	Encoder-Decoder + Spectral Norm
Loss Function	MSE + Perceptual Loss	Adversarial Loss + L1 Loss	Adversarial Loss + L1 Loss
Working Principle	Learns to restore images directly	Constructs new images through Generator–Discriminator	Similar to baseline GAN with stabilization (using Spectral Norm)
Average SSIM	0.82	0.87	0.91
Success Rate	68%	75%	88%
Processing Time	0.15 s	0.30 s	0.35 s
Strengths	- Faster than GANs - Suitable for less severe image Damage	- Generates more natural details - Produces sharper and more realistic structures	- High structural similarity - Reduces mode collapse problem - Achieves highest SSIM and success rate
Weaknesses	- Results may appear blurred - Ineffective with severe Damage	- Requires adversarial training - May encounter instability	- Slower processing time - Computational cost increases due to SN
Suitable for	Damage1–20%	Damage 20–40%	Damage 40–60% (Highest Efficiency)

TABLE V. PERFORMANCE COMPARISON ACROSS DIFFERENT LEVELS OF DAMAGE

Damage Level	GAN+SN	GAN (Baseline)	CNN (U-Net)
1–20%	95%	85%	80%
20–40%	88%	75%	70%
40–60%	81%	65%	54%
Average	88%	75%	68%

From Table IV and Table V, the performance comparison of the proposed GAN+SN model with other baseline methods is presented as follows: CNNs (U-Net): Utilizes a standard 5-layer U-Net architecture commonly applied in image-to-image translation tasks, trained with a combined loss function of Mean Squared Error (MSE) and Perceptual Loss. GANs (Baseline): Employs the same Generator and Discriminator architectures as the proposed model but without Spectral Normalization, trained with Adversarial Loss combined with L1 Loss to enhance image sharpness. GANs + Spectral Normalization (Proposed Model): Trained with Adversarial Loss, L1 Loss, and applies Spectral Normalization to the Generator. All models were trained and evaluated under the same conditions and dataset to ensure fairness in comparison.

The high scanning success rate (88%) demonstrates our model's effectiveness in reconstructing not just visual appearance but the functional data structure of QR codes, analogous to how shadow tomography reconstructs 3D information from limited 2D projections in microscopic imaging [13].

V. DISCUSSION

The comparison of the three main techniques for QR Code restoration showed that each method had different strengths and limitations. CNN (U-Net) was the fastest method, taking only 0.15 seconds per image. It achieved an SSIM of 0.82 and a scanning success rate of 68%, making it suitable for applications requiring high speed. However, it had limitations when the damage exceeded 20%, because although Perceptual Loss helped preserve the overall structure, the pixel-level accuracy enforced by Adversarial Loss and L1 Loss (especially when stabilized by SN in the GAN model) appeared to be more effective in reconstructing the fixed geometric patterns of QR Codes, which are essential for successful scanning. The baseline GAN provided a good balance between quality and speed, achieving an SSIM of 0.87 and a scanning success rate of 75%, taking 0.30 seconds per image. However, it sometimes faced training stability issues. The GAN + Spectral Normalization (GAN+SN), which was developed in this study, achieved the highest performance with an SSIM of 0.91 and a scanning success rate of 88%. It was able to handle damage up to 40–60%, although it required the longest processing time of 0.35 seconds per image. This illustrated the trade-off between speed and robustness to damage. GAN+SN was a more suitable choice when restoring heavily damaged QR Codes, providing better stability and reducing mode collapse. The choice of technique should depend on specific requirements: if high speed is prioritized, CNN (U-Net) is recommended for minor damage; for a good balance between quality and speed, the baseline GAN is appropriate; and if maximum performance is desired without much concern for processing time, GAN + Spectral Normalization provides the best results in terms of both accuracy and robustness to severe damage.

TABLE VI. COMPARISON OF PERFORMANCE WITH OTHER STUDIES BY CRITERIA

Criteria	GAN+SN	EHFP-GAN	SRGAN
Main Objective	Reconstruct missing regions	Reconstruct missing regions	Enhance image details
Performance (Success Rate)	Minor Damage (1–20%): 95% Moderate Damage	Mild Damage 95.35% [11]	Not directly evaluated [12]
Key Advantage	Resistant to severe Damage	Performs best with mild Damage	Most suitable for distorted QR Code
Main Limitation	Longer processing time compared to other methods	Performance significantly decreases with heavy Damage	Cannot recover information that is completely missing

From Table VI, when compared with the study by Zheng et al. (2023), which proposed the Edge-Enhanced Hierarchical Feature Pyramid GAN (EHFP-GAN), a two-stage architecture that focused on repairing edges first and then reconstructing the QR Code, it was observed that the EHFP-GAN approach was more complex [11]. In contrast, our study used a single-stage GAN architecture, but emphasized stabilizing the training

process by applying Spectral Normalization (SN) directly to the Generator layers. Our approach focused on improving the overall image reconstruction quality and the sharpness of the black-and-white structure, which is crucial for QR Codes, rather than separating the edge repair process.

In terms of performance, our GAN+SN model achieved an average scanning success rate of 88% for up to 60% damage, whereas the EHFP-GAN model reported a recognition rate of 95.35% for minor damage and dropped to 31.94% for severe damage [11]. Although EHFP-GAN achieved higher success rates on minor damage, our model demonstrated better robustness when facing high levels of damage (40 to 60%), which represented a particularly challenging scenario. Furthermore, compared with studies using Super-Resolution techniques such as SRGAN [12], which achieved a PSNR of 30.06 and SSIM of 0.936 for QR Code upscaling, our model, focusing on inpainting missing parts, still provided competitive results (PSNR 28.5, SSIM 0.91) and, importantly, verified effectiveness through the scanning success rate, the primary practical metric.

A key strength of our study was the use of Spectral Normalization, which reduced unstable training and mitigated Mode Collapse commonly encountered in GAN training. This allowed the model to better learn the distribution of complete QR Code data, resulting in improved reconstruction of fine details and sharp black-and-white modules. In contrast, more complex approaches like EHFP-GAN may require additional data and fine-tuning to achieve maximum performance. However, a limitation of our study was that processing time could be slightly longer than the CNN (U-Net) method (0.35 seconds compared to 0.15 seconds per image), although the improved accuracy was more significant for practical applications.

Although the model performed well on synthetic data, its effectiveness on real-world damage, such as stains, sun fading, or perspective distortions, remained a challenge that required further investigation.

VI. Conclusion

This study presented an approach for restoring damaged QR Codes using Generative Adversarial Networks (GANs) combined with Spectral Normalization (SN) to enhance training stability. Experiments on a dataset with simulated damage showed that the developed model could generate restored QR Codes with high sharpness and strong similarity to the original QR Codes, achieving an average PSNR of 28.5 dB and SSIM of 0.91. Moreover, the model significantly improved the scanning success rate of damaged QR Codes, reaching 88%, which outperformed the CNNs (U-Net) and basic GANs methods under the same experimental conditions.

The use of Spectral Normalization in the Generator enabled the model to effectively reconstruct complex details in the restored regions, reduce artifacts, and allowed the model to better learn the distribution of complete QR Code data. Although the proposed model required slightly longer processing time than other methods, the improved accuracy was highly significant for practical applications, particularly in scenarios where QR Codes were severely damaged.

For limitations and future work, this study had certain limitations that should be considered. The dataset used for training consisted of synthetic data, which may not cover the full variety of real-world damage. The model was also unable to handle complex damage, such as QR Code distortions or damage caused by reflections.

Overall, this study proved a promising approach for restoring damaged QR Codes using GANs combined with Spectral Normalization, which could be applied across various industries and scenarios where QR Codes play a critical role.

References

[1] H. Kato and K. T. Tan, "Pervasive 2D Barcodes for Camera Phone Applications," IEEE Pervasive Computing, vol. 6, no. 4, pp. 76-85, Dec. 2007, doi: 10.1109/MPRV.2007.80.

[2] L. Belussi and N. S. T. Hirata, "Fast QR Code Detection in Arbitrarily Acquired Images," in Proc. 24th SIBGRAPI Conf. on Graphics, Patterns and Images (SIBGRAPI), Maceió, Brazil, Aug. 2011, pp. 281-288, doi: 10.1109/SIBGRAPI.2011.16.

[3] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image Inpainting for Irregular Holes Using Partial Convolutions," in Computer Vision – ECCV 2018, Lecture Notes in Computer Science, vol. 11215, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer, 2018, pp. 89-105, doi: 10.1007/978-3-030-01252-6_6.

[4] ISO/IEC 18004:2024, Information technology — Automatic identification and data capture techniques — QR code bar code symbology specification, 4th ed., Aug. 2024. Geneva, Switzerland: International Organization for Standardization, 2024. [Online]. Available: https://standards.iteh.ai.

[5] S. Tiwari, "An Introduction to QR Code Technology," in Proc. 2016 Int. Conf. on Information Technology (ICIT), Bhubaneswar, India, Dec. 2016, pp. 39-44, doi: 10.1109/ICIT.2016.021.

[6] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative Image Inpainting with Contextual Attention," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 5505-5514, doi: 10.1109/CVPR.2018.00577.

[7] Y. Yu, F. Zhan, and S. Lu, "EdgeConnect: Generative image inpainting with adversarial edge learning," IEEE Trans. Image Process., vol. 30, pp. 2322–2332, 2021, doi: 10.1109/TIP.2021.3050689.

[8] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," in Proc. Int. Conf. on Learning Representations (ICLR), Vancouver, BC, Canada, Apr. 2018, arXiv:1802.05957, doi: 10.48550/arXiv.1802.05957.

[9] K. Song, N. Liu, Z. Gao, J. Zhang, G. Zhai and X. -P. Zhang, "Deep Restoration of Invisible QR Code from TPVM Display," 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), London, UK, 2020, pp. 1-6, doi: 10.1109/ICMEW46912.2020.9105961.

[10] Q. Ma, R. Chen, and W. Wang, "QR code images fast blind deblurring based on specific symbol prior and L2 norm minimization," in Proc. 2023 4th Int. Conf. on Computer Engineering and Application (ICCEA), Guangzhou, China, Apr. 2023, pp. 89-94,doi:10.1109/ICCEA58433.2023.10135499.

[11] J. Zheng, R. Zhao, Z. Lin, S. Liu, R. Zhu, Z. Zhang, Y. Fu, and J. Lu, "EHFP-GAN: Edge-Enhanced Hierarchical Feature Pyramid Network for Damaged QR Code Reconstruction," Mathematics, vol. 11, no. 20, p. 4349, Oct. 2023, doi: 10.3390/math11204349.

[12] Y. Sancar, “Reconstructing unreadable QR codes: a deep learning based super resolution strategy,” PeerJ Computer Science, vol. 2025, Art. no. e2841, Apr. 2025, doi: 10.7717/peerj-cs.2841.

[13] Mateev, Valentin, and Iliana Marinova. "Depth field reconstruction by shadow tomography for small digital microscope." AIP Conference Proceedings. Vol. 2505. No. 1. AIP Publishing LLC, 2022.

Recent Integrating Machine Learning and Malay-Arabic Lexical Mapping for Halal Food Classification

Noorrezam Yusop[1], Massila Kamalrudin[2], Nuridawati Mustafa[3], Tao Hai[4],
Mohd Nazrien Zaraini[5], Halimaton Hakimi[6], Siti Fairuz Nurr Sardikan[7]
Software Engineering Department, Fakulti Teknologi Maklumat dan Komunikasi,
Universiti Teknikal Malaysia Melaka, Durian Tunggal, Malaysia[1, 2, 3]
Information Technology Department-College of Engineering & IT, Ajman University, United Arab Emirates[4]
Department of Computer System and Communication-Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, Durian Tunggal, Malaysia[5]
Department of Computing, Universiti Teknologi PETRONAS, Perak, Malaysia[6]
Department of Agricultural and Biological Engineering Technology-Faculty of Plantation and Agrotechnology,
Universiti Teknologi MARA (UiTM) Cawangan Melaka Kampus Jasin, Melaka, Malaysia[7]

***Abstract*—The rapid growth of e-commerce has changed the way people engage with businesses, notably in the food industry. For the Muslim community, guaranteeing Halal conformity in digital transactions is critical. This study provides a comprehensive framework for improving Halal E-Commerce systems that include machine learning, pattern libraries, and multilingual support, specifically in Malay and Arabic. The study examines the role of pattern libraries in designing user-friendly interfaces, as well as lexical mapping strategies for enhancing Malay-Arabic translation accuracy. Natural language processing (NLP) and machine learning are combined to create an application that classifies food items into two categories: Halal or Haram. With an accuracy of 85%, a Random Forest classifier is trained on labeled datasets. Preparing the text, extracting features using TF-IDF, and evaluating the results using precision, recall, and F1-score are all steps in the classification process. To increase classification accuracy, a rule-based approach is also applied to conditional logic and keyword matching. By adjusting the parameters, the model is further improved, leading to strong performance. By taking into account the cultural and linguistic requirements of the Muslim community, multilingual support enhances accessibility and user confidence. The suggested method increases translation accuracy by employing lexical mapping at the word, phrase, and context levels. The paper also assesses several machine learning models, demonstrating that Random Forest outperforms the other methods examined. The findings contribute to the growth of Halal E-Commerce by outlining a systematic strategy to ensure compliance and usability. The proposed system can serve as a platform for future research into AI-driven Halal certification and digital marketplace optimization, blockchain with an e-Commerce framework.**

Keywords*—Malay-Arabic lexical mapping; natural language processing; machine learning; halal food classification; halal food e-commerce

I. Introduction

The rise of e-commerce has altered consumer behavior and modified global food supply chains, providing unprecedented access to a wide range of items [1]. During this digital transformation, the Muslim consumer group faces special hurdles, most notably assuring Halal compliance with food products. Halal, as a religious necessity, necessitates rigorous adherence to Islamic dietary requirements, and current Internet retail platforms typically fail to provide sufficient transparency and contextual information [2]. Product descriptions, certificates, and ingredients get increasingly complicated as they are written in multiple languages and cultural contexts. A fundamental issue in current Halal verification methods is their reliance on static certification markings, which may be out of date, regionally specific, or linguistically incomprehensible. Furthermore, the lack of real-time, automated, and linguistically inclusive verification methods severely impedes the development of culturally responsive digital food classification technologies [3]. Despite the growing emphasis on ethical consumption and religious inclusion, current platforms sometimes lack advanced computational techniques or multilingual semantic frameworks, both of which are required for effective Halal verification in cross-border e-commerce.

Simultaneously, modern usability requirements need systems that satisfy functional objectives while also providing a smooth and culturally adaptive user experience. According to user interface design and accessibility studies, pattern libraries are critical for creating consistent, intuitive interfaces that boost user confidence and interaction [4]. Multilingual support is equally vital, and Arabic remains the principal language of Islamic jurisprudence, whereas Malay is the predominant means of communication among Southeast Asian Muslim communities [5]. The purposeful use of various languages allows for semantic clarity and broad accessibility. Existing research has stressed Arabic's liturgical relevance as well as Malay's regional significance in the spread of Islamic knowledge [6]. However, there has been minimal research on developing lexical alignment between these languages for Halal food classification [7]. Given the linguistic discrepancies between ingredient labeling and certification forms, an intelligent system that bridges Arabic-Malay lexical mappings can drastically reduce

misinterpretation and improve classification accuracy. This gap in multilingual lexical integration within Halal certification systems is largely investigated in the literature.

Recent improvements in machine learning (ML) offer effective methods for automating difficult classification tasks. Machine learning algorithms can be trained to recognise non-Halal materials, assess multilingual product descriptions, and detect small linguistic cues suggesting Halal or Haram status [8][9]. However, their ability to classify Halal remains uncertain in the absence of culturally relevant linguistic information [10]. Thus, integrating machine learning with domain-specific Malay-Arabic lexical mapping improves context understanding, reduces false positives, and enables real-time, scalable Halal compliance solutions in food eCommerce platforms.

The paper describes an integrated system for automated Halal food classification in multilingual digital contexts based on machine learning and Malay-Arabic lexical mapping. The suggested technique bridges the research gap by incorporating cultural-linguistic information into data-driven models, thereby connecting technical capabilities to Muslim clients' religious and ethical requirements. This study adds to academic discourse and the development of inclusive, intelligent Halal eCommerce infrastructures by evaluating the existing state of the art, establishing a clear approach, and giving practical design solutions.

II. Literature Review

A. The Evolution of Halal e-Commerce

The rise of e-commerce has created new opportunities to connect specialist product categories with global audiences, particularly in faith-based marketplaces such as Halal food. Saleem et al. [1] argue that introducing Halal certification into digital platforms is more than simply a branding tool. It is critical for maintaining religious compliance and consumer trust. Nurillah [2] highlights the significance of cross-border transparency, traceability, and regulatory alignment in digital Halal ecosystems. Furthermore, Yener [3] claims that user-centric design, which incorporates culturally relevant factors such as dietary filters and religious certifications, may enhance trust and purchasing confidence among Muslim customers. These studies demonstrate a growing demand for digital solutions that not only permit compliance but also include Islamic values in the user experience (UX) design.

B. Role of Multilingual Support in e-Commerce

Language is a critical component of user interaction with digital services. Haque [5] argues that multilingual support improves accessibility greatly, particularly among linguistically heterogeneous user populations. For Halal systems, integrating Arabic (because of its religious significance) with Malay (in part due to its widespread adoption in Southeast Asia) assures not just language but also cultural resonance. According to Mabruroh and Khoiriyah [6], adopting these languages fosters diversity and matches the user experience with the expectations of observant Muslim customers. This concept contributes to the development of e-commerce systems that are not only linguistically translated but also semantically tied to religious and cultural significance.

C. Pattern Libraries in Software Development

Pattern libraries offer a systematic approach to interface design, promoting consistency, efficiency, and reuse in software development. Puteri and Widyanti [4] claim that using pattern libraries reduces cognitive work while improving system predictability for users. Haque et al. [5] developed culturally relevant pattern libraries for Islamic systems, such as prayer times, modest fashion filters, and Halal certification widgets. Additional research by [11] and [12] demonstrates that culturally localising pattern libraries improves usability while also generating emotional trust, which is an important factor in religious technology acceptance. These findings emphasise the necessity of culturally appropriate UI components in Halal-oriented systems.

D. Lexical Mapping

Lexical mapping functions as a linguistic bridge, enabling accurate translation and classification across languages. This is especially important in Halal systems, because the meaning of terminology varies depending on the cultural, gastronomic, or religious setting. Lexical mapping normally consists of four components: word-level, phrase-level, context-specific, and morphological mapping [13].

1) Word-level mapping: This approach is based on one-to-one lexical equivalence. For example, "makan" translates to "أكل" (food), and "rumah" to "بيت" (house), setting the framework for basic vocabulary alignment.

2) Phrase-level mapping: Fixed expressions, such as greetings and frequent idioms, are represented using units. For example, "Selamat pagi" becomes "صباح الخير" (Good morning), while "Terima kasih" maps to "شكراً" (Thank you), while keeping semantic intent.

3) Context-specific mapping: Words with several meanings require context-sensitive disambiguation. The Malay word "masak" can mean "ripe" (ناضج) in the context of fruit or "to cook" (طبخ) in culinary usage, highlighting the necessity for sophisticated disambiguation algorithms.

4) Morphological mapping: This includes root-word relationships and affixes. For instance, "makanan" (meal) is derived from the root "makan" and the suffix "-an," resulting in "طعام." Similarly, "berjalan" (walking) breaks down into "ber-" and "jalan," which corresponds to "يمشي" in Arabic. This morphological intelligence is necessary for developing robust natural language processing (NLP) systems.

III. Methodology

To design, develop, and test the recommended food categorisation model, a proof-of-concept application was built by merging Machine Learning (ML) techniques with Malay-Arabic lexical analysis via Natural Language Processing (NLP) [14]. This methodological framework is separated into three major stages: parameter design, analysis, and testing, as follows:

A. Parameter Design

The parameter design phase of the study focuses on overall process flow and component integration, as illustrated in Fig. 1. The system architecture uses machine learning and rule-based analysis to identify whether a food item is halal or not. The

program parses Malay-language food descriptions, possibly translating them into Arabic for better semantic understanding, before assessing the content using a hybrid classification technique.

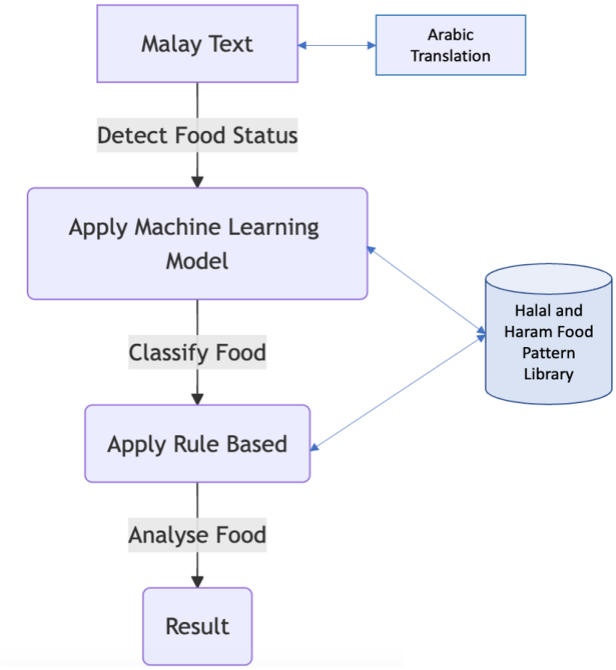


Fig. 1. Process activity for integrating machine learning and Malay-Arabic lexical mapping.

B. Parameter Selection

The Malay Arabic translation, machine learning model, and rule-based approach were considered when selecting parameters for this study. The particulars of parameter selection were explored in Fig. 2 as follows.

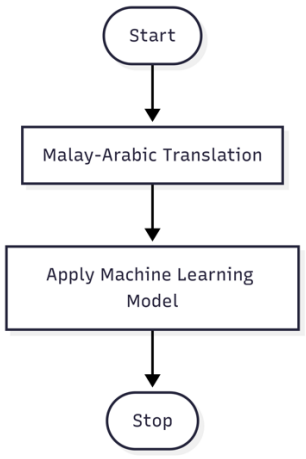


Fig. 2. Parameter selection.

1) Malay to Arabic translation: The system gets food-related text in Malay that needs to be evaluated. Consider "Daging ayam dengan alkohol" (alcoholised chicken meat). Finally, if necessary, the Malay text is translated into Arabic for better semantic understanding.

2) Apply machine learning model: The Machine Learning (ML) approach to categorising halal and haram foods consists of many key stages. Data collection and preprocessing begin with the compilation of a list of food names labeled as halal or haram according to Islamic dietary guidelines, followed by text preparation steps such as tokenisation, lowercasing, and special character removal, with words then transformed into numerical attributes using TF-IDF or word embeddings. Next, a Random Forest Classifier is used to increase accuracy by integrating numerous decision trees; this strategy takes advantage of classification and regression trees (CARTs) while reducing correlations between them [15].

During the model training and assessment phase, the dataset is separated into training (80%) and testing (20%) subsets, with the training data allowing the model to learn patterns and the testing data validating its predictive potential [16]. The model's performance is then evaluated using metrics such as accuracy, precision, recall, and F1-score: accuracy measures overall correctness, precision measures the proportion of correctly predicted positive cases, recall measures the ability to identify all actual positive cases, and the F1-score combines precision and recall to provide a balanced evaluation of performance [17].

3) Apply rule-based: A rule-based system uses established logical rules to classify and analyse data [18]. A rule-based approach for halal and haram food classification includes manually established rules, such as:

a) Keyword matching is the process of determining when a food item has halal or haram components by comparing its ingredients to a predetermined list of keywords. b) Simple criteria are employed to apply conditional logic, such as labelling a food as "haram" if it contains "babi" (pork) or "halal" if it contains "ayam" (chicken). This classification is based on the direct presence of specific terms. c) Natural language processing (NLP) is used to break down a sentence or meal description into individual words or tokens, which are then evaluated using predefined classification algorithms. This enables more effective handling of text input in many languages or unstructured formats. d) As the classification procedure progresses, decision trees are introduced.

C. Parameter Analysis

This section, as shown in Fig. 3, examines data collection or preparation, text pre-processing, random forest classifier training, and halal or haram phrase prediction, with the following details.

1) Data preparation: This structured dataset can serve as a foundation for machine learning models aimed at automatic halal food classification via natural language processing. Fig. 4 illustrates heatmaps comparing halal and haram classifications.

```
import pandas as pd

data = {"sentence": ["ayam","gelatin babi","susu","alkohol",
"daging lembu","babi","sayur-sayuran organik", "ikan", "babi",
"biawak", "telur", "Unta", "Kambing", "renet", "lemak babi",
"daging peha babi"],
```

"label": [1, 0, 1, 0, 1, 0, 1, 1,0,0,1,1,1,0,0,0]

1 = Halal, 0 = Haram}

df = pd.DataFrame(data)

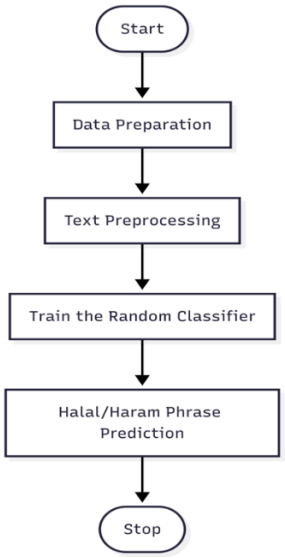


Fig. 3. Parameter analysis.

The collection contains 16 food-related items that have been classed as halal (permissible) or haram (forbidden) by Islamic dietary guidelines. Each item contains a food term (for example, "ayam", "gelatin babi", or "susu") and a binary label such as 1 for halal and 0 for haram. The data is stored in a pandas DataFrame, with column "sentence" containing the food's name or description and column "label" showing its classification. Eight of the 16 items are halal, including "ayam" (chicken), "susu" (milk), "daging lembu" (beef), and "ikan" (fish), while the remaining eight are haram, including "gelatin babi" (pork gelatin), "alkohol" (alcohol), "lemak babi" (pork fat), and "daging peha babi" (pork thigh meat). This structured dataset can serve as a foundation for machine learning models aimed at automatic halal food classification via natural language processing. Fig. 4 illustrates heatmaps comparing halal and haram classifications.

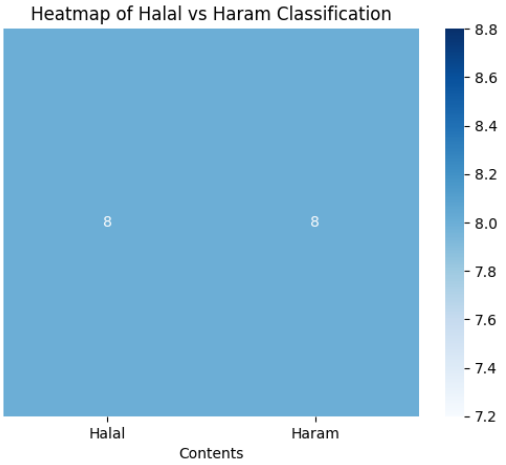


Fig. 4. Heatmaps of halal vs. Haram classification.

2) Text preprocessing: To prepare the data for machine learning, the TF-IDF vectorisation approach was used to convert text into numerical features. As a result, this Python technique is part of a pipeline that uses text classification to detect halal and haram food items. The data is separated into two categories: training (80%) and testing (20%). The following line explains how to utilise TF-IDF vectorisation to convert text into numerical attributes. Fig. 5 illustrates the steps for preparing text in code statements.

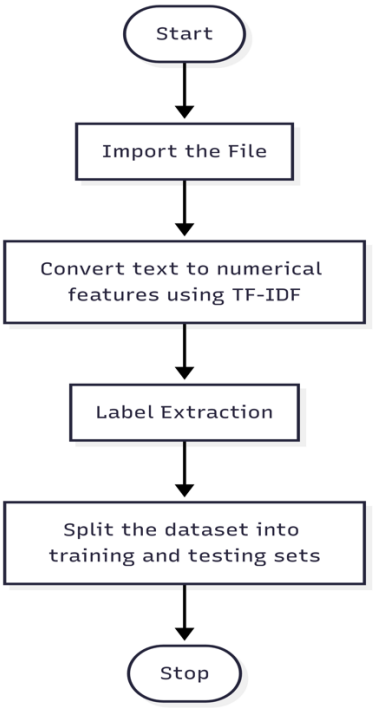


Fig. 5. Text preprocessing.

Based on Fig. 5, step 1, TfidfVectorizer from sklearn.feature_extraction.text turns text data into numerical representations that a machine learning model can process. The train_test_split from sklearn.model_selection partitions the dataset into training and testing sets for model evaluation. In Step 2, TF-IDF (Term Frequency-Inverse Document Frequency) is applied to the text data (df["sentence"]), turning words into numerical vectors representing their significance in the dataset. Stop_words='english' removes common words (such as "the" and "is") that do not contribute to classification. Step 3 extracts the label column, which contains binary categorisation values (1 for halal and 0 for haram). Finally, in Step 4, the dataset is partitioned into training (80%) and testing (20%) sets to assess model performance. This `random_state=42` ensures consistent splits across different runs, improving reproducibility.

3) Train the random forest classifier: The Random Forest model was trained and tested on pre-processed text data with scikit-learn for text categorisation. The purpose is to identify food items as halal or haram based on textual descriptions. Fig. 6 depicts the Random Forest Classifier training.

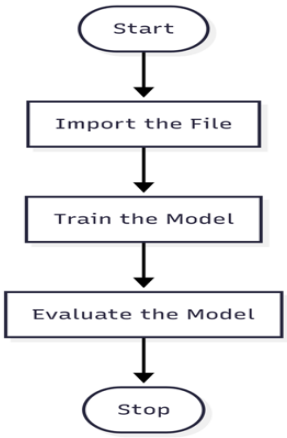


Fig. 6. Random forest classifier training.

Based on Fig. 6, step 1, RandomForestClassifier is imported from sklearn.ensemble. It is an ensemble learning technique that use numerous decision trees to boost classification accuracy. The accuracy_score from sklearn.metrics is used to assess the model's performance by determining the proportion of correctly categorised labels. In Step 2, n_estimators equal 100: The model generates 100 decision trees, with the final classification depending on the majority vote of these trees. The random_state is 42: Fixes the random seed to ensure reproducibility. The model is trained on the training dataset (X_train, y_train), learning patterns from text-based features extracted by TF-IDF. Finally, in Step 3, the trained model predicts the class labels (halal or haram) for an unknown test dataset (X_test). The model's accuracy is calculated by comparing predicted labels (y_pred) to actual labels (y_test). Accuracy is expressed as a decimal value (for example, 0.85 for 85%). Displays the accuracy score rounded to two decimal places.

4) Prediction on halal or haram sentences: The sentences or words are classified using the trained model based on the following Algorithm 1.

Algorithm 1: Sentences or words classification for prediction, Halal or Haram algorithm

```
Initialize
Compute
def classify_food (sentence):
predictions = []
sentence_transformed = vectorizer.transform([sentence])
predicts[count] = model.predict(sentence_transformed)
print (f"ML- predict: {predicts}"
count = count +1
if 0 in predicts:
return "HARAM"
else:
return "HALAL"
```

D. Parameter Testing

Parameter testing is necessary during implementation to optimise the Random Forest classifier's performance. Several important factors were meticulously altered and assessed to determine the best model design, as shown in Table I. First, the n_estimators parameter, which controls the number of decision trees in the forest, was tested for values ranging from [10, 50, 100, 200]. The trials found that increasing the number of estimators improved model stability and accuracy, with 100 trees providing a good balance of performance and processing efficiency. As a result, the final model used n_estimators = 100. The max_depth option, which determines the maximum depth of each tree, was tested with values of [5, 10, 15, None]. A depth of 15 was identified to appropriately reflect the data's complexity while preventing overfitting, which occurs when no limit (None) is used. The min_samples_split and min_samples_leaf variables were also useful for refining the model. Testing settings for min_samples_split as [2, 5, 10] and min_samples_leaf as [1, 2, 5] revealed that the default values of 2 for split and 1 for leaf resulted in better generalisation, keeping the model's flexibility while avoiding excessive splits that could lead to overfitting. The criterion parameter, which describes the function used to determine the quality of a split, was compared to both 'gini' and 'entropy'. The 'gini' index was chosen since it gave somewhat better performance and faster processing in this scenario. Finally, to ensure reproducibility, random_state=42 was kept constant throughout the tests. The final optimised model, which included these parameters, was trained using TF-IDF-transformed training data (X_train, y_train) and tested against test data (X_test, y_test). The printed end accuracy demonstrated that the chosen criteria resulted in a strong and accurate classification system for discriminating between "Halal" and "Haram" food items. The parameter setting is demonstrated in Algorithm 2 for RandomForestClassifier.

Algorithm 2: Parameter Testing for RandomForestClassifier

```
Initialize
Compute
optimized_rf = RandomForestClassifier(
n_estimators=100, max_depth=15, min_samples_split=2, min_samples_leaf=1, criterion='gini', random_state=42) optimized_rf.fit(X_train, y_train)
y_pred = optimized_rf.predict(X_test)
final_accuracy = accuracy_score(y_test, y_pred) print("Final Optimized Random Forest Accuracy:", final_accuracy)
```

TABLE I. TESTING N_ESTIMATORS (NUMBER OF TREES)

Parameter	Description	Value to Test
n_estimators	Number of decision trees in the forest	10, 50, 100, 200
max_depth	Maximum depth of each decision tree	5, 10, 15, None
min_samples_split	Minimum number of samples required to split a node	2, 5, 10
min_samples_leaf	Minimum samples required in a leaf node	1, 2, 5
criterion	Function to measure the quality of a split	'gini', 'entropy'

During implementation, the code is based on parameter selection and testing, and is designed to categorize food products for "Halal" or "Haram" using a set of phrases and labels. Then, the text input is converted into numerical characteristics for categorisation using machine learning (ML), a Random Forest classifier, and TF-IDF vectorisation. Algorithm 2 shows how the

ML model is trained using the sentences. After that, the accuracy is checked. Halal and haram foods and make recommendations for halal products is identified using a rule-based system. Google Translate to translate Malay API is used in a code, as well as machine learning and rule-based analysis to classify meals and provide halal alternatives. The application also shows how both methods can be used to categorise and recommend food items.

This study introduces a novel integration of Malay–Arabic lexical mapping into machine learning–based Halal food classification, a combination that has not been previously explored in existing literature. While prior studies have focused on single-language (Malay or English) Halal text classification, this research leverages bilingual semantic alignment between Malay and Arabic to enhance contextual understanding of religious terminology in ingredient lists. Furthermore, the application of the Random Forest classifier within this multilingual, culturally grounded framework represents a methodological advancement, as it enables higher interpretability and robustness compared to conventional models such as Naïve Bayes and SVM. The framework's strength lies in its ability to connect cultural linguistics and AI-driven food classification, resulting in a system better aligned with Islamic dietary principles and real-world e-commerce applications.

IV. Results

The study findings are separated into three categories which are TF-IDF Score analysis, Comparison of Performance analysis, and Precision, Recall, and F1-Score for Halal Food Malay-Arabic.

A. TF-IDF Score Analysis

Fig. 7 shows the TF-IDF heatmap, which visualises term importance across a variety of food-related texts. The heatmap illustrates the frequency and significance of specific words in the dataset, with darker hues (red) suggesting higher importance and lighter tints (blue) indicating lower relevance. This Figure is essential for understanding the distinction between halal and haram foods. Example keywords such as "babi" (pork) and "alkohol" (alcohol) have higher TF-IDF values in haram sentences, indicates that they are commonly employed to discriminate against prohibited products. Similarly, terms such as "ayam" (chicken) and "ikan" (fish) have high TF-IDF scores in halal-classified phrases, emphasising their importance in identifying permissible food products. The dataset's structure reveals that some terms are more distinctive, whilst others appear often in several sentences, resulting in lower TF-IDF scores. This graphic effectively underlines which words contribute the most to the categorisation process in a machine-learning model [19]. However, while the heatmap depicts word relevance, it has several disadvantages. Such as TF-IDF does not capture word context. Instead, it uses frequency-based weighting, which may not fully reflect real-world nuances in halal-haram classification [20]. For example, the term "daging" (flesh) can be halal or haram depending on the context, but TF-IDF cannot tell the difference. Furthermore, the heatmap's readability could be improved by sorting phrases by relevance or clustering related words. Combining TF-IDF with additional advanced NLP techniques, such as word embeddings or transformers, may increase classification accuracy by capturing deeper semantic connections [21]. Recent research indicates that deep learning models can outperform TF-IDF in halal food classification by analysing ingredient texts with contextual embeddings [22].

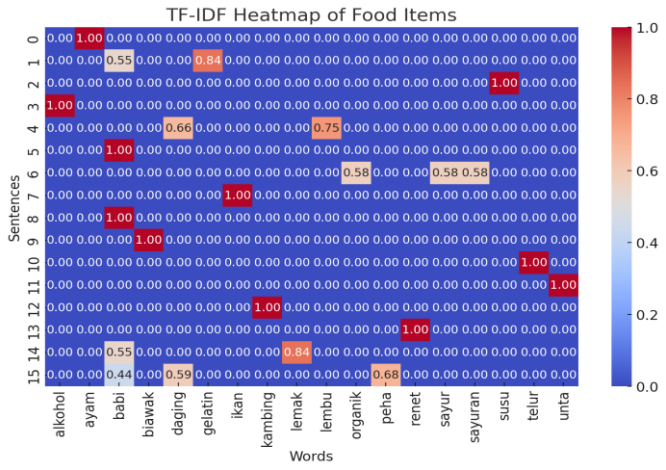


Fig. 7. TF-IDF of food items.

B. Performance Analysis

Fig. 8 compares Random Forest, Naïve Bayes, SVM, and Logistic Regression in terms of accuracy and training time.

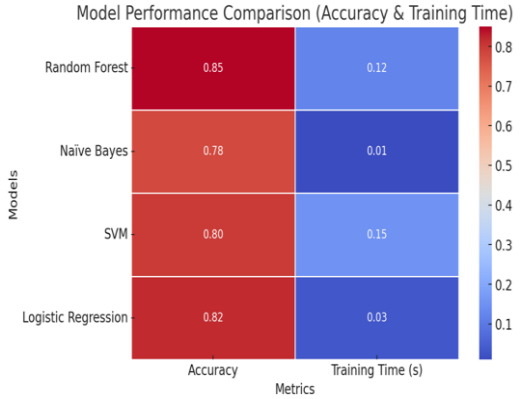


Fig. 8. Comparison of performance analysis.

According to Fig. 8, a comparison of machine learning models for halal-haram classification shows that Random Forest has the highest accuracy (85%), making it the most reliable alternative. However, it has a moderate training time of 0.12 seconds, which is longer than simpler models like Naive Bayes (78% accuracy, 0.01s training time). The fastest model, Naïve Bayes, is effective for quick classifications but lacks accuracy. Meanwhile, SVM (80% accuracy, 0.15s training time) performs admirably but takes the longest to learn. Logistic Regression (82% accuracy, 0.03s training time) achieves a balance of efficiency and interpretability, but it may struggle with complicated patterns.

C. Precision, Recall, and F1-Score for Halal Food Malay-Arabic

To assess the effectiveness of the proposed halal food classification system, particularly with the addition of Malay-Arabic lexical mapping, four machine learning models were

tested using three major classification metrics: precision, recall, and F1-score. Fig. 9 presents these variables as heatmap comparisons, which allows for an easier assessment of model performance across several evaluation criteria.

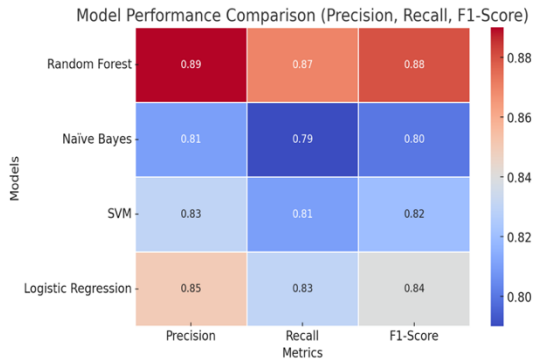


Fig. 9. Precision, recall, and FI-score.

Naive Bayes produces reasonable results (Precision: 0.81, Recall: 0.79, F1-Score: 0.80), indicating that, while it works adequately, it cannot generalise as well as Random Forest. Support Vector Machine (SVM) improves recall (0.81) and balance (F1-Score: 0.82) but falls short on precision (0.83), possibly due to its sensitivity to overlapping classes. Logistic Regression has a high precision (0.85) and a decent recall (0.83), yielding an F1-score of 0.84, indicating balanced but slightly poorer performance than Random Forest.

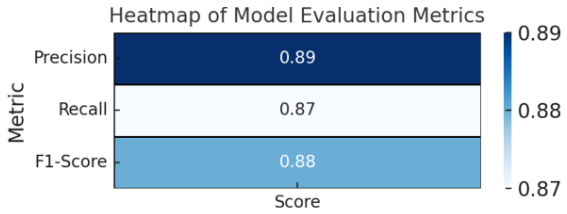


Fig. 10. Output of model evaluation metrics.

Fig. 10 demonstrates that the Random Forest model outperforms all three metrics. These high scores demonstrate that Random Forest properly identifies halal/haram classes (precision) while also catching the vast majority of actual positive examples (recall), yielding an exceptional balance (F1-score). This result indicates the model's stability and dependability in this classification job, especially when supplemented with lexical mapping between Malay and Arabic terminologies, which aids in the resolution of semantic ambiguities in food ingredients and naming traditions.

V. Discussions

Overall, Random Forest is the most reliable and accurate classification model, making it ideal for practical halal food identification. Naïve Bayes provides the fastest performance, while Logistic Regression balances accuracy and efficiency. SVM, while accurate, is slower and requires more computing power. The model chosen is based on the application's specific requirements, such as accuracy, speed, and interpretability [23][24]. These findings suggest that Random Forest is the most effective classifier for the Malay-Arabic halal food classification problem. Lexical mapping significantly enhances the semantic comprehension of ingredient lists by connecting culturally and religiously significant phrases in Malay and Arabic, such as recent studies by Khalilia et al. [25] demonstrate the creation of cross-lingual maps via a supra-lingual concept layer that can account for culturally unique terms that lack direct counterparts in other languages. This linguistic enrichment allows the model to discover minor distinctions that are critical in halal food assessment, hence boosting both the classification system's accuracy and cultural context [26]. Thus, combining machine learning and multilingual lexical analysis improves performance metrics while also aligning the system with real-world requirements in various Muslim-majority or multilingual regions.

VI. Conclusions and Future Work

This study emphasises the need to implement AI-powered technologies to improve Halal e-commerce platforms. Using machine learning classification models and lexical mapping methodologies, the system accurately detects Halal items while retaining their linguistic and cultural significance. The study found that combining pattern libraries, multilingual support, and advanced text processing significantly enhances usability, accessibility, and trust among Muslim consumers. The Random Forest model, achieving a maximum accuracy of 85% outperforming the baseline models by up to 10%, proved to be the most effective Halal–Haram classifier. The proposed framework lays the groundwork for future developments in Halal digital commerce, ensuring Islamic compliance while leveraging modern technological advancements. However, several limitations must be acknowledged. The dataset size used in this study is relatively limited, which may constrain the model's ability to generalize to broader food categories. Additionally, potential data bias in labeled samples and restricted language coverage, such as being focused mainly on Malay and Arabic that may affect the model's accuracy and cultural inclusiveness in wider contexts. Future research should therefore aim to expand multilingual datasets, address bias through balanced sampling, and improve NLP capabilities to enhance classification accuracy and contextual understanding. Moreover, integrating real-time certification verification with blockchain technologies, as well as an e-Commerce framework, could improve transparency and traceability in Halal authentication. Collaboration with Halal certification authorities and e-commerce providers will also be vital to standardize digital Halal verification procedures. Ultimately, this study contributes to the growing field of Halal e-commerce by paving the way for safer, more efficient, and culturally inclusive online markets, such as helping to eliminate doubtful Halal products and build greater consumer trust.

Acknowledgment

We would like thanks to Universiti Teknikal Malaysia Melaka (UTeM) and Ajman University for the grant number URMG-AJMAN/2024/FTMK/A00070 as well as Fakulti Teknologi Maklumat dan Komunikasi (FTMK) for their support.

References

[1] N. Saleem, A. Khalid, and S. Sadiq, "Impact of e-commerce on Consumer Behavior: A Study of Online Shopping Trend among Youth in Lahore

City," Ashwin Anokha Publications & Distributions, 2022. [Online]. Available: http://www.ashwinanokha.com/IJEB.php.

[2] S. L. Nurillah, "The Importance of Halal Certificates in Halal Food Using the Digital Platform," Int. J. of Law Dynamics Review, vol. 1, no. 2, Nov. 2023, pp. 99–111.

[3] D. Yener, "The Effects Of Halal Certification And Product Features On Consumer Behavior," Int. J. of Management Studies, vol. 29, 2022.

[4] R. Puteri and A. Widyanti, E-commerce of Islamic fashion product: usability and user acceptance. 2018.

[5] S. Haque, "Language Use and Islamic Practices in Multilingual Europe," 2025. [Online]. Available: https://www.cambridge.org/core.

[6] M. Mabruroh and R. Khoiriyah, "Islamization of Malay Language and its Role in the Development of Islam in Malaya," AJIS: Academic Journal of Islamic Studies, vol. 4, no. 1, Jul. 2019, p. 13.

[7] M. S. Mazli, "Deep learning-based halal food recognition," 2020.

[8] S. Tarannum, M. S. Jalal, and M. N. Huda, "HALALCheck: A Multi-Faceted Approach for Intelligent Halal Packaged Food Recognition and Analysis," in IEEE Access, vol. 12, 2024, pp. 28462-28474.

[9] A. M. Ahmad et al., "Ensuring halal food integrity: An overview of modern molecular and technological solutions," Food Biomacromolecules, vol. 2, 2025, pp. 5–22.

[10] B. Hutchinson, "Modeling the Sacred: Considerations when Using Religious Texts in Natural Language Processing," in Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 2024, pp. 1029–1043.

[11] K. Zurida Ishak, and A. Jaafar, "Cultural Dimensions of Malaysian Teenagers and Their Relationship with Interface Design", J. Theor. Appl. Inf. Technol., Vol. 90, No.2 (2016), pp. 176-186.

[12] S. M. Z. Samsuri, S. A. Ariffin, and N. S. Fathil, "Incorporating Cultural Design Elements in Mobile Applications Creative Industries in Malaysia: A Conceptual Study", J. ICT in Education, Vol. 8 No.2 (2021), pp. 110-117.

[13] A. François, "Lexical tectonics: Mapping structural change in patterns of lexification," De Gruyter, vol.41. no.1, 2022, pp.89-123.

[14] D. Alfter, "Complexity and Indecision: A Proof-of-Concept Exploration of Lexical Complexity and Lexical Semantic Change," 2024.

[15] Z. Sun, G. Wang, P. Li, H. Wang, M. Zhang, and X. Liang, "An improved random forest based on the classification accuracy and correlation measurement of decision trees," Expert Syst Appl, vol. 237, Mar. 2024.

[16] V. Roshan Joseph, "Optimal Ratio for Data Splitting," Statistical Analysis and Data Mining, Feb. 7, 2022.

[17] S. Sathyanarayanan, and B. Roopashri Tantri, "Confusion Matrix-Based Performance Evaluation Metrics," Afr. J. Biomed. Res., vol. 27, no. 4s, Nov. 2024, pp. 4023–4031.

[18] Z. Wang, J. R. Talburt, N. Wu, S. Dagtas, and M. N. Zozus, "A Rule-Based Data Quality Assessment System for Electronic Health Record Data," Appl Clin Inform, vol. 11, no. 4, Aug. 2020, pp. 622–634.

[19] F. Tobing, H. Fauzan, K. Simatupang, and R. Aulia, "The Role of Artificial Intelligence in the Halal Industry," Information Management and Business Review, vol. 16, no. 3(I)S, Sept. 2024, pp. 159-166.

[20] L. Kalyani, M. Jacintha, D. Kar, N. Roy, and V. K. Sharma, "A Comparative Analysis of Text Embeddings (TF-IDF, Word2Vec, FastText) for Machine Learning-Based Fake News Detection," Int. J. Eng. Res., vol. 11, no. 6, June 2024, pp. a760-a763.

[21] F. Rifaldy, Y. Sibaroni, and S. Suryani Prasetiyowati, "Effectiveness of Word2Vec and TF-IDF in Sentiment Classification on Online Investment Platforms Using Support Vector Machine," J. Penelitian dan Pembelajaran Informatika (JIPI), Mar. 2025.

[22] A. D. Bhargavi, "Comparative Study of Static and Contextual Text Vectorization for Sentiment Analysis," International Journal for Research in Applied Science & Engineering Technology, 2025.

[23] A. J. Wyner, S. J. Bleich, D. M. Olson, and B. Mease, "Explaining the Success of AdaBoost and Random Forests," Journal of Machine Learning Research, vol. 18, 2017, pp. 1-40.

[24] H. Ahmed Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," Babylonian J. of Machine Learning, vol. 2024, pp. 69-79, 2024.

[25] A. Khalilia et al., "Crowdsourcing lexical diversity: Mapping language-specific concepts into a supra-lingual semantic layer," ACM Trans. Asian Low-Resour. Lang. Inf. Process, vol. 37, no. 4, Article 111, Sept. 2024.

[26] N. Yusop, M. Kamalrudin, N. A. Moketar, and N. Mustafa, "The role of language in Malaysia's halal food industry: Trends, challenges, and future directions," Int. J. Res. Innov. Soc. Sci., vol. 9, no. 2, Feb. 2025, pp. 437–445.

Evaluating Head Pose Estimation for Assessing Visual Attention in Children with Special Needs During Robot-Assisted Therapy

Rusnani Yahya[1], Rozita Jailani[2,*], Nur Khalidah Zakaria[3], Fazah Akhtar Hanapiah[4]
Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM), Malaysia [1, 2, 3]
Faculty of Medicine, Universiti Teknologi MARA, Sungai Buloh, Malaysia [4]
Center for Medical Electronic Technology, Politeknik Sultan Salahuddin Abdul Aziz Shah, Shah Alam, Malaysia [1]

***Abstract*—This study investigates the application of head pose estimation (HPE) to assess visual attention in children with special needs (CwSN) during robot-assisted therapy sessions, focusing on its effectiveness and the attention patterns exhibited by these children. CwSN often faces unique challenges, such as sensory processing difficulties or delayed cognitive processing. Age and therapy duration also influenced attention levels, with younger children generally exhibiting shorter attention spans than older participants. Additionally, familiarity with technology, such as prior screen time at home, positively impacted engagement during robot-assisted therapy. An experimental study was conducted with 30 children aged 2 to 7 years, including those with autism spectrum disorder (ASD), speech delay (SD), and attention-deficit/hyperactivity disorder (ADHD). Using an integrated camera, head movements were tracked to analyse forward-facing head direction as an indicator of attention. The system achieved an overall accuracy of 82% and an average attention percentage of 65%, highlighting that visual attention varies significantly based on the type of disability, age, and therapy duration. The integration of the robot enhanced visual engagement across all groups, fostering improved interaction and attention. These findings emphasise the importance of tailoring robot-assisted therapy (RAT) to the specific needs and attention patterns of children with different disabilities, ages, and therapy histories, underscoring the potential of assistive robotics to optimise therapeutic outcomes in special education settings. This research highlights the potential of personalised RAT to improve social, cognitive, and motor skills. It offers evidence-based strategies for integrating assistive robotics into special education and therapeutic settings for CwSN.**

***Keywords*—Head pose estimation; visual attention; robot-assisted therapy; children with special needs**

I. Introduction

Visual attention assessment is critical for understanding and enhancing therapeutic outcomes for CwSN. These children often exhibit atypical visual attention patterns, which impact their social interactions, communication skills, and learning processes [1] [2]. Accurate assessment of visual attention provides valuable insights into cognitive and behavioural states, enabling therapists to tailor interventions that promote engagement and skill development. Nevertheless, traditional observation-based assessments predominantly remain subjective and demonstrate variability across different sessions, thereby limiting the accuracy and reliability of therapeutic evaluations.

To address these limitations, there is an increasing need for an AI-driven methodology capable of objectively quantifying attention through head pose estimation (HPE). The understanding and supervision of visual attention using such computational techniques are essential for improving engagement in therapeutic and educational settings, in alignment with the global movement towards intelligent and inclusive educational technologies. The HPE, specifically analysing head direction patterns, is significant in assessing attention in CwSN [1]. Monitoring head movements helps therapists understand where a child is focusing their attention during therapy sessions [2], enabling real-time feedback and adaptive strategies that foster improved social and cognitive outcomes [3].

However, quantifying and accurately assessing head direction in CwSN presents significant challenges. These children may struggle to maintain a consistent gaze or display rapid and unpredictable head movements, adding complexity to understanding head direction patterns unique to this group[4]. Gaining more profound insights into these patterns is crucial for tailoring interventions to support their developmental needs [5].

Despite significant advancements in head pose estimation, few studies have utilised these methodologies to evaluate visual attention among children during real-world therapeutic sessions. Current models primarily focus on adult datasets or controlled laboratory environments, thus limiting their applicability in educational and clinical contexts. This research aims to address this limitation by developing and validating a real-time head pose estimation framework designed to assess visual attention within the scope of robot-assisted therapy.

This study evaluates head direction patterns in CwSN during RAT sessions. Using an integrated camera system to track head movements toward a service robot, we aim to identify distinct attention patterns associated with CwSN. The novelty of this work lies in the integration of a hybrid HPE algorithm into real-world therapy sessions, as well as the multi-factor analysis of attention patterns across disability type, age, and therapy duration. These contributions advance the field of RAT by providing real-time, evidence-based insights that can

* Corresponding author.

guide the design of more personalised engagement strategies and improve therapeutic outcomes for CwSN.

The remainder of this study is organised as follows: Section II reviews related work on head pose estimation and visual attention assessment. Section III describes the methodology, including the proposed framework, data collection, and analysis procedures. Sections IV and V presents the results and discussion, while Section VI concludes the study with implications, limitations, and directions for future research.

II. Literature Review

A. Head Pose Estimation Techniques

Head pose estimation (HPE) is pivotal in interpreting visual attention and social interactions, especially among CwSN. Various methods have been developed to estimate head pose, broadly categorised into feature-based, appearance-based, and hybrid approaches. Feature-based methods rely on detecting facial landmarks such as the eyes, nose, and mouth to infer head orientation [6]. The algorithms and techniques associated with these methods can be categorised into classical approaches, machine learning-based models, deep learning algorithms [16], and hybrid techniques. Classical algorithms, such as Active Shape Models (ASM) and Active Appearance Models (AAM), focus on statistical shape modelling and texture analysis [7][8].

At the same time, machine learning-based approaches utilise classifiers like Support Vector Machines (SVM) and Random Forests for head pose prediction [9][10][11]. Deep learning algorithms, including Convolutional Neural Networks (CNNs) and heatmap regression methods, leverage advanced architectures to enhance accuracy and robustness in landmark detection and pose estimation [12]. Appearance-based methods employ machine learning techniques to analyse pixel intensity patterns across the face without explicitly detecting facial features. CNNs have been widely adopted in this category due to their robustness in handling variations in lighting and facial expressions. These methods have advanced with the development of deep learning, enabling more accurate and efficient head pose estimation.

Significant advancements in HPE techniques for assessing visual attention in CwSN have been achieved. Deep learning approaches, including CNNs and transformer-based models, have gained prominence due to their ability to handle the complex and variable head movement characteristic of CwSN populations. These models have been optimised for real-time performance, enabling seamless integration into interactive therapy sessions without causing delays or disruptions [13][14].

Hybrid techniques, such as the Perspective-n-Point (PnP) algorithm, integrate 2D facial landmarks with 3D head models for precise orientation estimation [6][10]. These advancements contribute significantly to the field of HPE, with applications spanning robotics, healthcare, and human-computer interaction. While effective, these methods can face challenges with occlusions and variations in facial expressions, which are common in real-world settings.

Researchers have developed non-intrusive methods using robot cameras and AI-based frameworks such as dlib and MediaPipe to capture head movements without requiring children to wear any devices. Techniques that leverage facial landmark detection and 3D modelling have improved the accuracy of HPE, even in uncontrolled environments with varying lighting conditions and occlusions. These advancements are essential for creating responsive therapeutic tools that can adapt to the unique behaviours of CwSN.

In the context of CwSN, real-time and adaptive HPE techniques are crucial. CwSN may exhibit rapid or atypical movements, making robust and flexible estimation methods essential. Recent developments have focused on enhancing these methods to increase accessibility and practicality for use in therapeutic settings, ensuring minimal intrusion while maintaining maximum accuracy.

B. Visual Attention in Children with Special Needs

Children with special needs (CwSN) often face challenges in maintaining attention, particularly in social contexts, where they may focus less on facial features and eye regions than typically developing peers, instead directing their attention toward objects or patterns. This atypical visual attention can hinder the development of social and communication skills, making it a critical area of interest for researchers. Studies employing eye-tracking technology have revealed that these children spend significantly [15] less time engaging with social stimuli, such as people, and more time attending to non-social elements within their environment. These findings suggest that interventions to redirect their visual attention toward socially relevant stimuli could enhance social skill acquisition [17]. Additionally, difficulties with joint attention and the ability to share a focus on an object or activity with another person are common and significantly impact language development. Targeted strategies to improve joint attention could, therefore, play a vital role in supporting the developmental needs of these children [18].

Recent advancements in research and technology have paved the way for innovative interventions tailored to the unique requirements of CwSN. Visual dysfunction, prevalent in this population, necessitates diagnostic approaches that extend beyond traditional visual acuity assessments. [19]. For example, integrating Augmented Reality (AR) in learning environments has demonstrated the potential to improve visual attention by providing multi-sensory experiences that actively engage children [20]. Additionally, tools such as 3D-printed toys have been shown to improve attention spans and fine motor skills in children with autism [21]. Environmental factors also play a critical role, with studies indicating that exposure to greenery, such as classroom windows overlooking natural landscapes, can positively impact sustained attention levels. Such findings highlight the importance of creating enriched environments and using adaptive tools to foster visual and cognitive engagement in CwSN [22] [23].

Educational strategies and structured interventions further enhance visual attention and related skills. Visual media, such as flashcards, have been shown to improve memory and recognition in autistic children [24]. Art provides an effective medium for visual expression among CwSN [25]. These strategies align with cognitive load theory, which emphasises tailoring instructional methods to manage cognitive demands

and optimise learning outcomes. By understanding how CwSN allocate their visual attention, therapists and educators can design targeted interventions that enhance engagement with social stimuli and promote meaningful improvements in communication, social interaction, and overall learning.

C. Robot-Assisted Therapy

Robot-assisted therapy (RAT) has emerged as a promising tool for supporting CwSN, particularly enhancing their social, emotional, and cognitive skills. Recent advancements have focused on designing robots to address specific challenges faced by children with autism. For instance, robot functionalities tailored for Applied Behaviour Analysis (ABA) therapy demonstrate the ability to stimulate cognitive skills while adapting to individual limitations [26]. A dual-cycle therapy model emphasising the integration of therapists as teleoperators, which enhances the effectiveness of robot-mediated interactions with neurodivergent children [27].

Social assistive robots have also been shown to facilitate emotional expression and secure social interactions in therapy sessions. These robots provide a judgment-free environment, significantly improving engagement and therapeutic outcomes in children with autism spectrum disorders (ASD) [28]. Furthermore, the use of social humanoid robots as mediators in interventions emphasises their potential to promote communication skills and interaction between children, teachers, and therapists [29]. The importance of advancing robot design lies in better accommodating the needs of CwSN.

In addition to fostering social and emotional growth, robot-assisted therapy is being explored in broader contexts, such as mental health interventions and physical skill development. The potential of robots to engage adolescents in therapeutic settings, with implications for similar applications in CwSN [30]. Integrating interactive elements in robot-mediated therapy has proven effective in enhancing engagement, indicating that these tools hold significant promise for complementing traditional therapy methods across diverse therapeutic contexts.

In summary, previous research has demonstrated promising methodologies for head pose estimation and attention tracking; however, most are confined to controlled environments or adult populations. Furthermore, the integration of these techniques into robot-assisted therapy remains insufficiently explored. To address these limitations, this study presents a hybrid framework combining Dlib and MediaPipe for robust head pose estimation and employs it to assess visual attention among children with special needs.

III. Methodology

The study involved an in-situ experiment at the Kizzu Kids Rehabilitation and Enrichment Centre, focusing on CwSN. Service robots were integrated into the participants' standard therapy routines in these sessions. Data was collected during these sessions, capturing video recordings, images, and tabular data in CSV files. These multimodal data sources were utilised to investigate visual attention patterns, particularly analysing how the children directed their gaze and engaged with their surroundings during therapy.

The collected data were used to evaluate a real-time system for measuring visual attention. Section III A describes the experimental procedure, outlining the data acquisition process and the specific formats for the captured data. In Section III B, the study introduces system architecture, detailing its components and incorporating a widely adopted HPE algorithm. This algorithm enables the assessment of gaze direction and engagement levels, thereby facilitating an analysis of the children's visual attention during RAT sessions.

A. Data Gathering from the In-situ Experiment

This section describes an experiment in which therapy sessions were recorded to create a database for machine learning research. These sessions involved the use of service robots to support CwSN, including those diagnosed with ASD, ADHD, and SD. The robot facilitated app-based activities that complemented therapy sessions by providing structured tasks across multiple domains, including communication (AutiSpark), social-emotional (LogicLike), cognitive (Khan Academy Kids), and motor domains (ABC Kids). These apps functioned primarily as mediators, ensuring the children remain engaged with the robot interface. At the same time, the core focus of the study was the integration of the hybrid head pose estimation (HPE) system. During these sessions, the robot's integrated camera and HPE workflow (dlib, MediaPipe, and PnP) simultaneously recorded and analysed head movements to quantify visual attention in real-time.

Thirty CwSN underwent the RAT to train skills across multiple domains, highlighting the potential of integrating technology and robotics into therapeutic intervention.

1) Participants: Thirty children diagnosed with ASD, ADHD, and SD were selected who are currently receiving treatment at the Kizzu Kids Rehabilitation and Enrichment Centre (Malaysia), a specialised institution for the rehabilitation of CwSN.

Ethical approval was obtained from the ethics committees of Universiti Teknologi MARA and Kizzu Kids Rehabilitation and Enrichment Centre. All the parents signed consent forms before their children were included in the study. Children were free to leave the experiment at any time and were always supported by a professional educator, other than the researcher.

2) The robot therapist: Temi 3 service robot: The robot that led the RAT was the Temi 3 service robot (see Fig. 1), a telepresence robot and a trustworthy autonomous personal assistant focused on high-quality video. By combining artificial intelligence and autonomous navigation, it can recognise and follow people on demand, memorise predefined locations, and navigate effortlessly in various settings, including hotels, restaurants, shops, businesses, educational institutions, healthcare facilities, and more.

This study utilised the default settings and standard equipment of the Temi 3 robot unless otherwise specified. The Temi 3 has multiple advanced cameras designed to enhance its functionalities. It features a 13-megapixel high-resolution camera with autofocus, capable of recording 1080p video at 30 frames per second (FPS), a 120-degree field of view (FOV), a

5-element lens, and a hybrid infrared (IR) filter. Additionally, it includes a wide-angle 13-megapixel camera for remote navigation, offering a 95-degree FOV and 1080p video recording at 30 FPS. Complementing these is a Time-of-Flight (TOF) depth camera, which operates at 30 FPS with a 90-degree FOV and an effective range of up to 5 meters, enabling depth perception essential for navigation and object detection. Among its software features, Temi 3 includes face detection and tracking capabilities, which are employed in *in-situ* experiments to guide the robot toward the child during interactions.

Fig. 1. The Temi 3 Robot. The camera used for recording the child's head pose is the one on top.

3) Protocol for the In-situ experiments: To evaluate the children's attention, this study focused on participants who were already enrolled in therapy sessions of varying durations and had been diagnosed by specialists before commencing the program. In this research, pre-existing app-based learning modules were downloaded and integrated into a robot interface to facilitate interaction between the children and the robot during therapy sessions. These modules were carefully selected to align with the therapeutic needs of the children, incorporating tasks adapted to their developmental levels. These tasks are designed to observe how children visually engage with the robot and respond to visual stimuli presented through the robot's interface.

The robot served as a mediator during the therapy sessions, integrating pre-existing apps and a camera system to facilitate interaction and observation. It was included in the children's daily activities and identified through a specific "visual schedule". A visual schedule communicates the sequence of upcoming activities or events using objects, photographs, icons, words, or a combination of these supports. During each session, the children engaged in tasks presented through the robot, with the apps delivering visual stimuli and the camera capturing data to observe and analyse the children's visual attention.

To facilitate interaction, the robot-led sessions were conducted in the same room where the children typically received their therapy, ensuring a familiar and comfortable environment. The robot was positioned in front of the child and initially placed at a distance of at least 0.5 meters. Children were allowed to adjust their seating position to enhance comfort. Each session focused on a single activity to evaluate visual attention, with tasks presented in a randomised sequence to minimise repetitive or predictable patterns. The robot introduced each activity using clear and straightforward verbal instructions and visual prompts to engage the children in the tasks effectively (see Fig. 2).

Fig. 2. An example of a child–robot interaction during the therapeutic session. A therapist was always present nearby to support the child.

A therapist familiar with the children's daily treatment routines was present during the sessions to provide security and support. The therapist offered positive reinforcement through verbal cues, such as "good" or "right", and, in some instances, physical reinforcement, such as a gentle touch or pat. These reinforcement strategies were individualised to meet each child's needs and address their unique behaviours.

Before the main sessions, an introductory session was conducted to familiarise the children with the robot and minimise potential novelty effects. During this preliminary session, the robot was introduced in a non-therapeutic context for approximately ten minutes, allowing the children to acclimate to its presence in a relaxed and informal manner.

4) Video recording and annotation

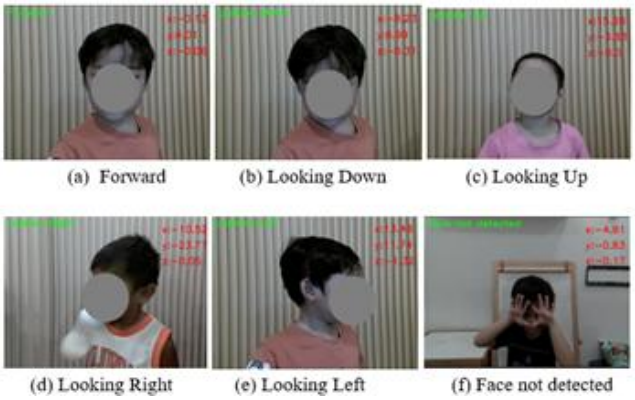


Fig. 3. The frame is extracted from four videos recorded by the robot's camera (the child interacts with the robot, while the system records and annotates the head position in real-time).

Fig. 3 above showcases a series of frames extracted from videos recorded by the Temi 3 robot's camera, capturing various head poses of a child during the interaction. Each frame represents a different head position, annotated in real-time by the system. The poses include: a) Forward, where the child looks directly at the camera; b) Looking Down, where the child tilts their head downward; c) Looking Up, with the child tilting their head upward; d) Looking Right, with the head turned to the right; e) Looking Left, where the child turns their head to the left; and f) Face Not Detected, where the face is obscured or not visible to the system. These annotated frames demonstrate the system's capability to record and categorise head positions during live interactions, essential for analysing visual attention and engagement.

B. Head Direction Detection Algorithm

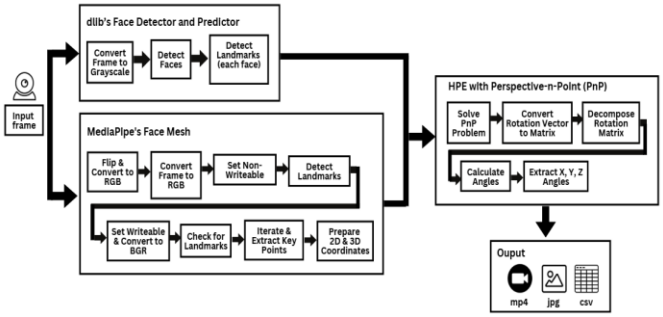


Fig. 4. Comprehensive workflow of the head pose estimation.

Fig. 4 illustrates a comprehensive workflow for HPE, integrating dlib's face detector and MediaPipe Face Mesh as complementary modules for robust facial landmark detection and analysis. The hybrid approach, which combines Dlib's 68-point landmark predictor with MediaPipe Face Mesh, was selected to leverage Dlib's geometric precision alongside the real-time tracking stability provided by MediaPipe. This integration enhances accuracy across diverse lighting conditions and head movements, which are commonly encountered during therapeutic sessions involving children.

To estimate the three-dimensional head orientation, the Perspective-n-Point (PnP) algorithm was employed to assess the three-dimensional head pose due to its efficiency and appropriateness for monocular camera input, thereby ensuring compatibility with standard RGB webcams integrated into the robotic platform.

Input video frames from a camera are preprocessed through two distinct pathways. The first pathway utilises dlib, where frames are converted to grayscale, facial regions are detected, and 68-point facial landmarks are extracted. This pathway ensures accurate detection and tracking of facial features under various lighting and environmental conditions. In the second pathway, MediaPipe's Face Mesh processes the input frames by converting them to RGB format and, if necessary, flipping them for consistency. The frames are then set to a non-writable state to optimise computational performance before detecting facial landmarks. These landmarks undergo iterative refinement to extract key points, which are further processed to generate 2D and 3D coordinate mappings. This dual-pathway approach enhances the system's robustness, enabling it to accommodate diverse input formats while maintaining precision across various scenarios.

The extracted key points from both pathways are input into the HPE module, which employs the Perspective-n-Point (PnP) algorithm for pose computation. This algorithm calculates head orientation by solving for rotation and translation vectors, which are then decomposed to derive X, Y, and Z rotation angles. These angles represent the real-time spatial orientation of the head, enabling precise monitoring and tracking of head movements. The integration of the dlib and MediaPipe modules ensures that the system achieves a balance between computational efficiency and detection accuracy.

Finally, the outputs from the workflow are presented in multiple formats to support diverse analytical needs. Annotated videos are generated in MP4 format, individual frame images are saved as JPG files, and a CSV file is produced to log timestamps, rotation angles, and head directions data. These multimodal outputs provide a comprehensive dataset for analysing head pose dynamics, thereby supporting applications in visual attention assessment and behavioural research. The modular design of this workflow ensures adaptability and effectiveness for real-time analysis in therapeutic environments.

C. Performance Measures

The forward-facing head pose measure was calculated to evaluate engagement and focus of the participants during therapy sessions, providing a quantitative metric for assessing visual attention. This was achieved by determining the total number of instances where the forward head pose was detected in real-time using head pose estimation algorithms. The forward-facing pose, which indicates that the child is looking directly at the task or stimuli, was analysed explicitly as a key performance measure of attention.

To calculate the frequency of forward-facing head poses as a measure of attention, the following formula was used:

$$\text{Frequency of Forward Head Pose (\%)} = \frac{T}{(T+F)} \text{ x } 100 \quad (1)$$

In this formula, T represents the total number of instances where the system correctly detected the forward head pose and verified its accuracy based on video frame analysis, indicating the participant's direct focus on the task or stimulus. F represents the total number of instances where the system detected a non-forward head pose (left, right, up, or down). Still, upon reviewing the video frames, these instances were verified to be correct forward head poses. This formula provides a normalised percentage of forward-facing head poses, enabling a consistent evaluation of visual attention during the therapy session. This formula provided a normalised percentage of forward-facing poses, enabling a consistent and comparable assessment across participants and sessions.

The forward-facing head pose was considered a reliable indicator of attention, reflecting how participants maintained their focus on the task or robotic intervention during therapy. Higher frequencies of forward-facing poses were interpreted as higher levels of engagement and attention, highlighting the ability of the participant to stay visually connected to the activity.

This analysis offered valuable insights into engagement patterns, emphasising forward-facing head poses as a direct measure of visual attention. Using this metric, the study demonstrated the importance of monitoring forward-facing head poses as a performance measure, particularly for children with special needs. This approach provides an objective and real-time method for evaluating engagement, offering actionable data to optimise therapy strategies and improve outcomes.

IV. Results

This section presents the experimental outcomes obtained from the head pose estimation framework, including detection accuracy, frame rate performance, and attention classification metrics across participant groups.

A. Accuracy of Head Direction Detections

The head direction detection algorithm accurately identified and categorised the five predefined head orientations: forward, left, right, up, and down. Validation against the automated annotated dataset showed an overall accuracy of 82% across all participants and sessions. Fig. 5 shows the accuracy percentage for each participant.

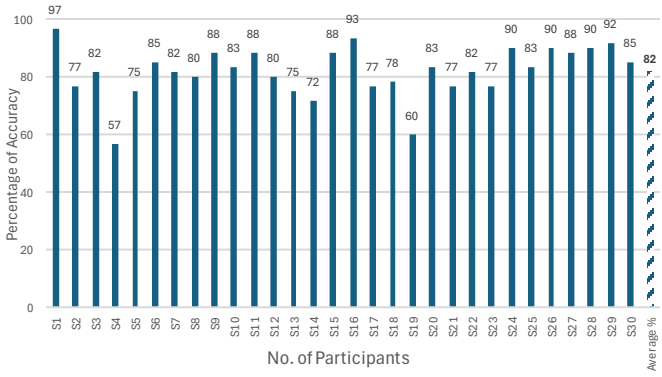


Fig. 5. The overall accuracy percentage for each participant.

Among the 30 participants, a clear contrast in accuracy rates was observed, with S4 and S19 recording the lowest rates and S1 and S16 achieving the highest. S4 achieved 57% accuracy, while S19 recorded 60%. Both participants are relatively inexperienced in therapy, with S4 having attended only a single session and S19 completing four sessions. This limited exposure may have contributed to their lower performance, as they are likely still developing familiarity and engagement with the tasks. Additionally, their young ages, 2 years for S4 and 4 years for S19, may reflect less developed cognitive and attentional abilities than older participants, further influencing their accuracy levels.

In contrast, S1 and S16 demonstrated the highest accuracy rates, with S1 attaining 97% and S16 achieving 93%. Their superior performance can be attributed to their extensive therapeutic exposure and structured intervention programs. S1, a 4-year-old female, has undergone eight one-to-one therapy sessions for personalised guidance and consistent engagement. Similarly, S16, a 6-year-old male, has participated in an intensive Early Intervention Program (EIP), attending sessions five days per week for 24 months. This rigorous and targeted approach likely enhanced their cognitive and attentional abilities, as well as their familiarity with therapeutic tasks, enabling them to outperform their peers.

B. Percentages of Attention Among Children with Special Needs

The attention percentages of CwSN were analysed to understand their engagement levels during interactions. Focusing on the proportion of time they maintained a forward head pose, a key indicator of attention, the analysis aimed to provide insights into their focus patterns. This measure is particularly significant in assessing how children with special needs respond to stimuli in structured settings, highlighting their ability to stay engaged and attentive over time.

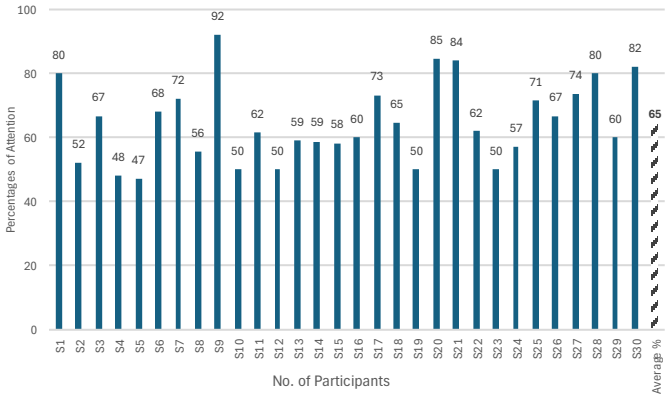


Fig. 6. The overall visual attention percentage for each participant.

Fig. 6 illustrates the attention percentage for each participant during the study, showcasing a wide range of engagement levels. The average attention percentage is 65%, with notable variations at both extremes. These differences highlight individual factors that influence the participant's ability to remain attentive during therapy sessions, including therapy duration, the severity of autism, age, and home environment.

The participant with the highest percentage of attention is S9, achieving an impressive 92%. S9 is a 5-year-old child with mild autism who has undergone 10 months of therapy and regularly experiences screen time at home. These factors may contribute to their higher attention level, as mild autism and longer therapy duration will likely enhance their ability to focus and engage with the task. Additionally, familiarity with screen-based interactions at home could make the experimental setup more engaging and comfortable for them.

In contrast, the participant with the lowest attention percentage is S5, recording just 47%. S5 is a 2-year-old child with ASD who has only undergone 9 months of therapy and does not have screen time at home. These factors might explain their lower attention level, as young ages and limited therapy duration can pose more significant challenges in maintaining focus. Furthermore, the lack of exposure to screens may make the experimental setup unfamiliar or less engaging for S5. These findings emphasise the importance of tailoring interventions to individual needs and considering factors such as therapy duration, different disabilities, and home environment when analysing engagement levels.

The differences in attention percentages highlight the importance of accounting for individual variability when interpreting these results. Participants with higher attention percentages, such as S9, may reflect the effectiveness of the experimental design in fostering engagement, particularly for children who benefit from longer therapy and familiar environments. Conversely, participants with lower attention levels, like S4, point to areas for improvement, such as minimising distractions or adapting activities better to suit the needs of children with more severe challenges. Understanding these patterns provides valuable insights for designing more inclusive and effective interventions, ensuring all participants can maximise their engagement during similar studies or therapeutic settings.

C. Type of Disability and Attention Pattern Analysis

The analysis of head direction among thirty children with special needs revealed distinct patterns among the three diagnostic groups: Autism Spectrum Disorder (ASD), Sensory Disorder (SD), and Attention Deficit Hyperactivity Disorder (ADHD).

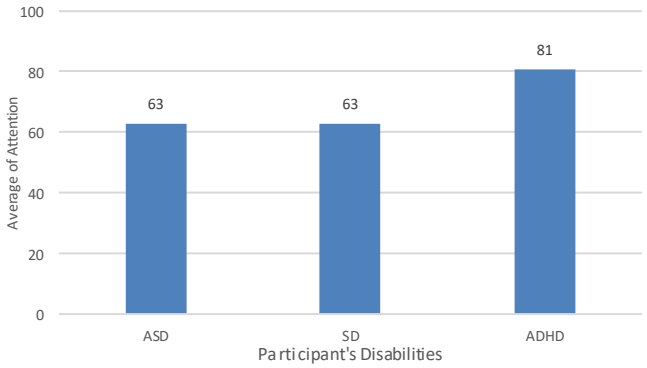


Fig. 7. The average of visual attention for different types of disability.

Fig. 7 illustrates the average percentages of visual attention across three categories of disabilities. The results indicate that children with ADHD (S20, S21, and S27) exhibit the highest average attention percentage at 81%, followed by those with SD at 63% and ASD at 63%. This suggests that the type of disability significantly influences attention levels, with children with ADHD demonstrating comparatively higher focus during activities.

From observations and discussions with a therapist during the therapy sessions, it was noted that children with ADHD tend to exhibit hyperfocus on tasks that are highly stimulating or engaging. These behaviours align with their preference for dynamic and interactive activities that effectively capture their attention. Additionally, shorter, highly interactive sessions are efficient in maintaining focus. These structured and engaging therapy strategies likely contribute to the higher average attention percentages observed in children with ADHD compared to other disability groups. Children with ASD and SD might face different challenges that affect attention, such as sensory processing difficulties or delayed cognitive processing, which could contribute to slightly lower average attention levels compared to children with ADHD. These findings may highlight a strength within the ADHD population with an ability to hyperfocus under the right conditions, which could be leveraged to improve therapeutic and educational outcomes.

D. Age-Based Variations in Attention Percentages Analysis

Visual attention plays a critical role in understanding the engagement and focus levels of CwSN during therapy sessions. By analysing attention percentages across different age groups, developmental patterns influencing their ability to concentrate on tasks can be identified. Below is a graph (see Fig. 8) illustrating the average attention percentages for children aged 2 to 7, which highlights the relationship between age and visual attention capabilities.

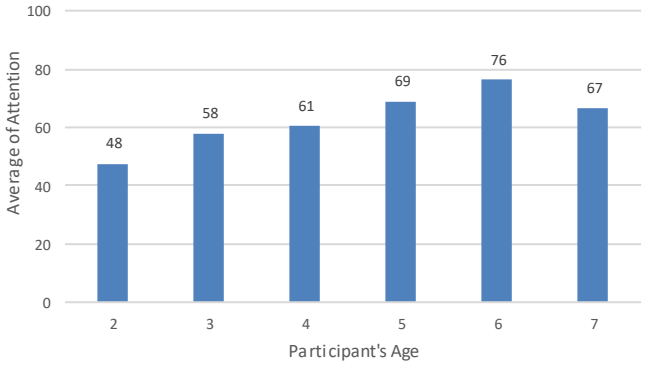


Fig. 8. The average visual attention for different ages.

The results reveal a noticeable trend where attention percentages increase with age, peaking at 76% for 6-year-olds before slightly declining to 67% for 7-year-olds. Younger children, particularly 2-year-olds (S4 and S5), exhibit the lowest attention levels at 48%, indicating developmental differences in focus and engagement during therapy sessions. This pattern suggests that as children grow older, their ability to maintain attention improves, likely due to increased cognitive development and maturity. However, the slight decrease in the oldest age group (7 years) may reflect varying individual differences or a plateau in attention development. These findings demonstrate that attention levels tend to align with developmental maturity as children grow older, highlighting the natural progression of focus and engagement with increasing age.

E. Impact of Therapy Duration on Attention

Therapy duration plays a significant role in influencing the attention levels of CwSN. Analysing attention percentages across different therapy durations provides valuable insights into how the length of therapy impacts focus and engagement over time. It is important to note that the children in this study were interacting with a robot for the first time, and their unfamiliarity with the robotic environment may have influenced their attention levels. The novelty of the robot could have initially captured their curiosity and engagement.

Fig. 9 illustrates the average percentages of visual attention for CwSN across varying therapy durations, ranging from 6 months to 36 months. The data reveal that children in the 6-month therapy group exhibited the highest attention levels at 67%. However, as therapy duration increased to 12, 24, and 36 months, attention percentages slightly decreased and stabilised

at around 63-64%. This pattern suggests that shorter therapy durations may initially result in heightened attention, potentially due to the novelty effect and intensive engagement during the early stages of therapy. In contrast, stabilising attention levels over longer therapy durations may indicate that children adapt to the therapy routines or experience a plateau in attentional improvement. These findings underscore the need to revise and adjust therapeutic strategies over time to sustain engagement and optimise long-term outcomes.

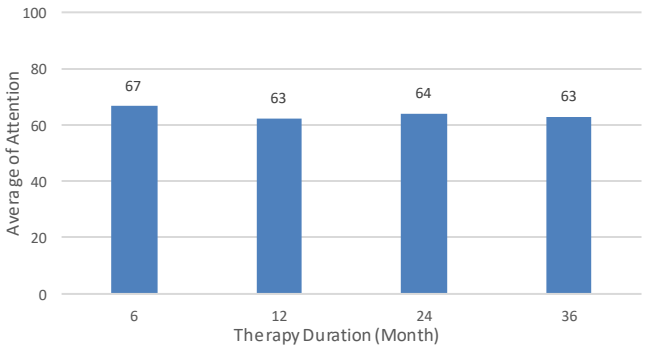


Fig. 9. The percentages of visual attention for different therapy durations.

V. Discussion

The results indicate that the proposed framework effectively identifies head orientation with high accuracy and stability. This section discusses the implications of these findings, compares them with previous studies, and highlights potential applications as well as existing limitations.

The outcomes of this study demonstrate the feasibility of integrating a head pose estimation algorithm within the context of robot-assisted therapy (RAT) to evaluate visual attention in children with special needs (CwSN). The system achieved an overall head orientation detection accuracy of 82%, alongside an average attention rate of 65%, indicating consistent performance across diverse participant groups and experimental scenarios. These results align with previous studies on real-time head pose estimation, which reported accuracy rates ranging from 78% to 85% under controlled conditions [31][32][33]. This consistency supports the robustness of the hybrid Dlib–MediaPipe–PnP methodology, demonstrating its resilience in the dynamic and unpredictable environment of therapy sessions.

Notably, the findings also reveal considerable variability in the results, suggesting that factors such as age, therapy experience, and type of disability have a substantial impact on engagement levels. This variation highlights the importance of considering developmental and contextual variables when interpreting performance metrics, as children with greater therapy exposure and familiarity with structured activities tend to achieve higher accuracy and attention scores.

Beyond technical validation, the study emphasises the broader implications of integrating robotics into therapeutic practice. RAT not only offers an objective and real-time method for attention monitoring but also facilitates opportunities for tailoring interventions to individual needs. The observed differences across various diagnostic groups and therapy durations indicate that a "one-size-fits-all" strategy may be inadequate, and that customised approaches are imperative to maintain long-term engagement.

These findings contribute to the growing body of evidence supporting artificial intelligence (AI) and robotics as effective tools for enhancing therapeutic precision, inclusivity, and personalisation. By integrating computer vision with behavioural analytics, the proposed framework demonstrates the potential of AI-driven systems to objectively quantify visual attention, thereby providing therapists with actionable data to support informed clinical decision-making. This integration represents a significant step toward data-informed, child-centred therapy environments that foster measurable developmental outcomes and elevate the overall quality of interventions for children with special needs.

Although the proposed framework demonstrates robust performance, several limitations should be acknowledged. The sample size was relatively small, and the controlled experimental setting may not fully reflect the variability and complexity of real-world therapeutic environments. Furthermore, external factors such as inconsistent lighting conditions and occasional occlusion intermittently influenced detection accuracy, suggesting the need for further refinement and validation in more diverse and naturalistic contexts.

Notwithstanding these limitations, the findings hold significant implications for therapeutic practice and special education. By enabling objective monitoring of attention, the framework allows therapists to tailor interventions and track engagement patterns over time, thereby enhancing therapeutic precision and potentially improving developmental outcomes for children with special needs.

In summary, this discussion highlights that the proposed hybrid head pose estimation framework demonstrates strong accuracy, real-time performance, and adaptability across diverse therapeutic contexts. The variations observed among participants underscore the importance of personalised and data-driven therapy strategies, wherein AI-based tools serve to complement rather than replace human expertise. By transforming head orientation data into meaningful attention metrics, this study bridges computational analysis with behavioural interpretation, enabling a more precise understanding of engagement in children with special needs. These insights establish a solid foundation for advancing intelligent robot-assisted systems and inform the development of future research directions in data-driven therapeutic interventions.

VI. Conclusion

The findings of this study highlight the significance of head direction patterns in assessing visual attention among children with special needs (CwSN), with variations in engagement, such as the high focus of S9 versus the challenges faced by S5, underscoring the influence of disability severity, therapy duration, and age. These results have important implications for robot-assisted therapy (RAT), as head direction data can inform the design of therapeutic activities better aligned with individual needs. Nonetheless, the study is limited by its relatively small sample size, variability in participant profiles,

and occasional inaccuracies in head pose classification. These limitations could be mitigated through larger, multicenter studies, the inclusion of multimodal behavioural measures (e.g., eye tracking, body posture), and further optimisation of the hybrid HPE algorithm to improve robustness in real-world conditions. Overall, the study emphasises the potential of head direction analysis in improving visual attention assessment and optimising RAT, offering a promising direction for enhancing therapeutic outcomes for CwSN.

This study presents a validated head pose estimation framework designed to quantify visual attention in children with special needs during robot-assisted therapy. Theoretically, it advances the understanding of how AI-based visual cues can represent attentional behaviour in non-verbal or minimally responsive participants. Practically, the system provides a non-intrusive solution that enhances engagement assessment within therapeutic and inclusive learning settings. Future research will involve larger participant groups, integration with eye-gaze and emotion detection technologies, and deployment in real therapeutic environments to validate scalability and adaptability.

Overall, the study establishes a solid foundation for the integration of AI-driven attention assessment in therapy, bridging computational modelling with behavioural science. By promoting more personalised, data-informed, and child-centred approaches, this research contributes to the evolving landscape of intelligent robot-assisted therapy, paving the way for future innovations that enhance developmental outcomes for CwSN.

ACKNOWLEDGMENT

This work was funded by the Strategic Research Partnership (SRP) grant (100-RMC 5/3/SRP INT (047/2023)) and (100-RMC 5/3/SRP INT (048/2023)). The authors would like to thank the Institute of Research Management Centre (RMC) and the Faculty of Engineering, Universiti Teknologi MARA (UiTM) Shah Alam, Malaysia, for the financial support, instrumentation, and experimental facilities provided. This study was undertaken during the study leave of the first author under the 2022 Federal Training (HLP) scholarship scheme awarded by the MOHE Malaysia.

REFERENCES

[1] M. Chita-Tegmark and M. Scheutz, "Assistive Robots for the Social Management of Health: A Framework for Robot Design and Human–Robot Interaction Research," Int. J. Soc. Robot., vol. 13, no. 2, pp. 197–217, 2021, doi: 10.1007/s12369-020-00634-z.

[2] A. Kouroupa, K. R. Laws, K. Irvine, S. E. Mengoni, A. Baird, and S. Sharma, "The use of social robots with children and young people on the autism spectrum: A systematic review and meta-analysis," PLoS One, vol. 17, no. 6 June, pp. 1–25, 2022, doi: 10.1371/journal.pone.0269800.

[3] T. Singh, "Attention Span Prediction Using Head-Pose Estimation with Deep Neural Networks," IEEE Access, vol. 9, pp. 142632–142643, 2021, doi: 10.1109/ACCESS.2021.3120098.

[4] Y. F. Kaul, K. Rosander, C. von Hofsten, K. Strand Brodd, G. Holmström, and L. Hellström-Westas, "Visual tracking at 4 months in preterm infants predicts 6.5-year cognition and attention," Pediatr. Res., vol. 92, no. 4, pp. 1082–1089, 2022, doi: 10.1038/s41390-021-01895-8.

[5] F. Yuan, E. Klavon, Z. Liu, R. P. Lopez, and X. Zhao, "A Systematic Review of Robotic Rehabilitation for Cognitive Training," Front. Robot. AI, vol. 8, no. May, pp. 1–24, 2021, doi: 10.3389/frobt.2021.605715.

[6] D. A. Kaliukhovich et al., "Social attention to activities in children and adults with autism spectrum disorder: effects of context and age," Mol. Autism, vol. 11, no. 1, pp. 1–14, 2020, doi: 10.1186/s13229-020-00388-5.

[7] Z. Shi, T. R. Groechel, S. Jain, K. Chima, O. Rudovic, and M. J. Matarić, "Toward Personalized Affect-Aware Socially Assistive Robot Tutors for Long-Term Interventions with Children with Autism," ACM Trans. Human-Robot Interact., vol. 11, no. 4, 2022, doi: 10.1145/3526111.

[8] Q. Wu and L. Zhang, "A Real-Time Multi-Task Learning System for Joint Detection of Face, Facial Landmark and Head Pose," pp. 1–12, 2023, [Online]. Available: http://arxiv.org/abs/2309.11773

[9] N. Van Nam and N. T. N. Quyen, "FLASH: Facial Landmark Detection Using Active Shape Model and Heatmap Regression BT - Industrial Networks and Intelligent Systems," N.-S. Vo and H.-A. Tran, Eds., Cham: Springer Nature Switzerland, 2023, pp. 170–182.

[10] G. Tzimiropoulos, J. Alabort-I-Medina, S. Zafeiriou, and M. Pantic, "Generic active appearance models revisited," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 7726 LNCS, no. PART 3, pp. 650–663, 2013, doi: 10.1007/978-3-642-37431-9_50.

[11] H. Su, W. Qi, J. Chen, C. Yang, J. Sandoval, and M. A. Laribi, "Recent advancements in multimodal human–robot interaction," Front. Neurorobot., vol. 17, 2023, doi: 10.3389/fnbot.2023.1084000.

[12] N. Efthymiou et al., "ChildBot: Multi-robot perception and interaction with children," Rob. Auton. Syst., vol. 150, 2022, doi: 10.1016/j.robot.2021.103975.

[13] S. Malek and S. Rossi, "Head pose estimation using facial-landmarks classification for children rehabilitation games," Pattern Recognit. Lett., vol. 152, pp. 406–412, 2021, doi: 10.1016/j.patrec.2021.11.002.

[14] A. Singh, K. Raj, T. Kumar, S. Verma, and A. M. Roy, "Deep Learning-Based Cost-Effective and Responsive Robot for Autism Treatment," Drones, vol. 7, no. 2, pp. 1–18, 2023, doi: 10.3390/drones7020081.

[15] N. Wedasingha, P. Samarasinghe, L. Senevirathna, M. Papandrea, A. Puiatti, and D. Rankin, "Automated anomalous child repetitive head movement identification through transformer networks," Phys. Eng. Sci. Med., vol. 46, no. 4, pp. 1427–1445, 2023, doi: 10.1007/s13246-023-01309-5.

[16] F. M. Talaat, "Real-time facial emotion recognition system among children with autism based on deep learning and IoT," Neural Comput. Appl., vol. 35, no. 17, pp. 12717–12728, 2023, doi: 10.1007/s00521-023-08372-9.

[17] S. Congiu, G. Doneddu, and R. Fadda, "Attention toward Social and Non-Social Stimuli in Preschool Children with Autism Spectrum Disorder: A Paired Preference Eye-Tracking Study," Int. J. Environ. Res. Public Health, vol. 21, no. 4, 2024, doi: 10.3390/ijerph21040421.

[18] C. F. Pinto, H. Mohan, R. Shenoy, V. Guddattu, and S. Tiwari, "The Effect of Parent-Mediated Joint Attention Intervention on Joint Attention and Language Skills in Children with Autism Spectrum Disorder - A Systematic Review," Child Fam. Behav. Ther., vol. 46, no. 3, pp. 272–297, 2024, doi: 10.1080/07317107.2024.2338741.

[19] M. Musleh, A. Green, A. Mankowska, C. Viner, and R. Pilling, "Spectrum of Visual Dysfunction Detected by a Novel Testing Protocol Within a Special School Eye Care Service," Br. Ir. Orthopt. J., vol. 20, no. 1, pp. 219–225, 2024, doi: 10.22599/bioj.391.

[20] G. Nie et al., "An Immersive Computer-Mediated Caregiver-Child Interaction System for Young Children with Autism Spectrum Disorder," IEEE Trans. Neural Syst. Rehabil. Eng., vol. 29, pp. 884–893, 2021, doi: 10.1109/TNSRE.2021.3077480.

[21] N. M. S. -, S. N. A. M. B. -, L. Z. Q. -, R. T. Y. S. -, and N. S. A. R. -, "Fine Motor Training using Conventional Tools, 3D-printed Toys and Digital Platform," Int. J. Multidiscip. Res., vol. 5, no. 6, pp. 1–8, 2023, doi: 10.36948/ijfmr.2023.v05i06.9934.

[22] K. D. C. M. Nushara and A. Hettiarachchi, "Impact of Greenery in the Window View on Visual Sustained Attention with Special Reference to Classrooms of Children with Down Syndrome," J. Real Estate Stud., vol. 20, no. 2, 2023, doi: 10.31357/jres.v20i2.6749.

[23] S. Bektaş and Z. G. Ercan, "A Study of Visual Motor Skills of Children With Special Needs," Eur. J. Educ. Stud., vol. 10, no. 8, pp. 341–359, 2023, doi: 10.46827/ejes.v10i8.4930.

[24] Y. Y. Subaihah, S. Alfinuha, M. Hasanah, and I. Indrawati, "Improved Ability to Recognize Letters Through Flashcard Media With Maximization Memory for Children With Autistic Special Needs," J. Univ. Muhammadiyah Gresik Eng. Soc. Sci. Heal. Int. Conf., vol. 2, no. 1, p. 95, 2023, doi: 10.30587/umgeshic.v2i1.5128.

[25] A. S. Dewi and E. Sugiarto, "Visual Expression of Children with Special Needs in SLB Muhammadyaah Surya Gemilang in the Image of the Picture," Gondang J. Seni dan Budaya, vol. 7, no. 1, p. 127, 2023, doi: 10.24114/gondang.v7i1.38420.

[26] T. S. F. Valenca, E. A. N. Carvalho, E. O. Freire, L. Molina, and J. G. N. De Carvalho Filho, "Specification of Robots for the Treatment of Autistic Children," Proc. - 2023 Lat. Am. Robot. Symp. 2023 Brazilian Symp. Robot. 2023 Work. Robot. Educ. LARS/SBR/WRE 2023, pp. 672–677, 2023, doi: 10.1109/LARS/SBR/WRE59448.2023.10333018.

[27] S. Elbeleidy, T. Mott, D. Liu, E. Do, E. Reddy, and T. Williams, "Beyond the Session: Centering Teleoperators in Socially Assistive Robot-Child Interactions Reveals the Bigger Picture," Proc. ACM Human-Computer Interact., vol. 7, no. CSCW2, 2023, doi: 10.1145/3610175.

[28] Polyxeni Ntaountaki, Georgia Lorentzou, Andriana Lykothanasi, Panagiota Anagnostopoulou, Vasiliki Alexandropoulou, and Agathi Stathopoulou, "Robotics for Autistic Children," Int. J. Sci. Res. Arch., vol. 9, no. 2, pp. 548–559, 2023, doi: 10.30574/ijsra.2023.9.2.0556.

[29] A. Puglisi et al., "Social Humanoid Robots for Children with Autism Spectrum Disorders: A Review of Modalities, Indications, and Pitfalls," Children, vol. 9, no. 7, 2022, doi: 10.3390/children9070953.

[30] P. Alves-Oliveira, T. Budhiraja, S. So, R. Karim, E. Björling, and M. Cakmak, "Robot-mediated interventions for youth mental health," Des. Heal., vol. 6, no. 2, pp. 138–162, May 2022, doi: 10.1080/24735132.2022.2101825.

[31] H. Lee, B. Oh, and S. C. Kim, "Recognition of Forward Head Posture Through 3D Human Pose Estimation With a Graph Convolutional Network: Development and Feasibility Study," JMIR Form. Res., vol. 8, pp. 1–12, 2024, doi: 10.2196/55476.

[32] S. Zhou, W. Zhang, Y. Liu, X. Chen, and H. Liu, "Real-Time Driver Attention Detection in Complex Driving Environments via Binocular Depth Compensation and Multi-Source Temporal Bidirectional Long Short-Term Memory Network," Sensors, vol. 25, no. 17, pp. 1–19, 2025, doi: 10.3390/s25175548.

[33] M. C. Ye and J. J. Ding, "Real-Time Head Orientation and Eye-Tracking Algorithm Using Adaptive Feature Extraction and Refinement Mechanisms †," *Eng. Proc.*, vol. 92, no. 1, 2025, doi: 10.3390/engproc2025092043.

A Novel Taxonomy for Human Activity Recognition Based on a Systematic Analysis of Public UAV Datasets

Sumaya Abdulrahman Altuwairqi*, Salma Kammoun Jarraya
Department of Computer Science-Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—In recent decades, unmanned aerial vehicles (UAVs) have become widely utilized for many real-world applications, including surveillance, crowd management, and threat detection, providing a new perspective to recognize human behaviors. However, current UAV-based video datasets adopt categorization schemes that rely on broad and inconsistent categories relative to real-world aerial contexts. To address this knowledge gap, this study proposes a novel human activity categorization framework derived from a comprehensive systematic analysis study of ten publicly available UAV-based human action recognition (HAR) datasets, incorporating a variety of environmental situations and human behaviors. By reconciling inconsistent categories and finer activities, this taxonomy serves as a standard framework for UAV-based HAR research. The proposed categorization framework is validated by comparing it with other existing frameworks on the publicly benchmarked Drone-Action dataset, outperforming them by 97% across four metrics. Our contribution aims to develop the foundation for further experimental validation and provide a guide for researchers interested in developing accurate and context-aware surveillance systems.**

***Keywords*—Human action recognition; UAV videos; surveillance systems; categorization framework**

I. Introduction

Unmanned aerial vehicles (UAVs), also known as drones, are distinguished by the lightweight integration of communication and sensor devices and their automation, as well as their power to provide wide-view coverage through rapid navigation across different locations. This makes them highly useful for a wide range of civilian and military applications, including reconnaissance, surveillance, scientific research, sports, and entertainment. UAVs have seen widespread use worldwide over recent years due to rapid technological evolution and their applicability in various systems. At present, the security and military sectors, rather than the industrial and commercial sectors, dominate the use of UAVs, which can be attributed to the increasing demand for traffic surveillance and monitoring [1-3], threat detection [4], search and rescue (SAR) [5], and intelligence, surveillance, and reconnaissance (ISR). Consequently, processing, interpreting, analyzing, and understanding UAV-based images and videos has become essential and highly desired.

One of the UAV exploitations that has garnered substantial attention is human action recognition (HAR), which is one of the promising fields of computer vision aimed at designing and building systems and methods that facilitate the automatic understanding, interpretation, and identification of various human activities in different contexts or environments captured in videos. HAR forms the cornerstone of a wide range of real-world applications, including visual surveillance systems, dense crowd management [6], smart healthcare, sports analytics [7], and the recognition of pedestrian actions for autonomous vehicles. UAV-based HAR visuals provide an unprecedented perspective for analyzing human behavior in various scenarios, making it a crucial topic for understanding and studying the human activities and behaviors that aid in building robust automated HAR systems suitable for UAV-based visual challenges.

Nowadays, several UAV-based datasets have been developed to capture diverse categories of human behaviors in real-world environments. Although these datasets have contributed significantly to the field, most of the recent studies [8-20] applied the categorization frameworks on these datasets that rely on full/partial single criteria of human actions and behaviors, resulting in broad, fine-grained, inconsistent, or overlapping categories. This issue may result in inaccuracies, as some critical human behaviors are ignored by being categorized inconsistently compared to their real-world context. For instance, in the Okutama-Action dataset [21], “pushing” may be categorized as someone moving a shovel's container forward, while the same classification is applied to someone forcefully pushing another person in another dataset. For the UAV-Human dataset [22], "walking" is fragmented into many categories based on the UAV's angle of view. These discrepancies and overlaps can confuse and impede the development of robust HAR-based approaches [8, 14, 23] due to their categorization lacking the ability to determine human activity levels and a reliance on human action properties, including kinematics, intentionality, and context dependency, which are essential for providing accurate and comprehensive categorization. Consequently, this highlights the urgent requirement for a systematic categorization scheme that reflects the subtle dissimilarities of human activities in aerial contexts.

To address the aforementioned gaps, this study proposes a novel human activity categorization framework designed

* Corresponding author.

specifically for UAV-based videos. The proposed categorization framework is based on three theoretical criteria of human behavioral and motion: 1) human action properties; 2) human behavioral characteristics; and 3) human activity levels. This framework derives from an in-depth, comprehensive analysis of ten publicly available UAV-based HAR datasets, involving a collection of human activities and contexts, and examines their challenges and complexities. Through systematically identifying similar and distinct human action categories across these datasets, we propose a unified yet precise taxonomy that considers the detail required for context-aware recognition. This taxonomy aims to serve as a foundational framework, allowing researchers to adopt consistent human behavior categories and provide insights into how to design and develop HAR-based systems that adapt to UAV challenges. To the best of our knowledge, this is the first study to introduce an in-depth study of human activity categories across multiple UAV-based HAR datasets and propose a unified taxonomy under aerial situations. The main contributions of this study include the following:

- Building a novel category for video-based human action recognition.
- Presenting a systematic study of ten public UAV-based HAR datasets, providing insightful analysis and discussion of their human activities and challenges, and defining various taxonomies, including levels, video categories, label schemes, environments, and factors of the UAV footage.
- Identifying a new behavior on the Okutama-Action dataset based on the human activity levels.
- Developing a unified and accurate human activities categorization framework founded on three criteria, compatible with UAV scenarios.
- Evaluating the performance of the proposed categorization by comparing it to the current categorization on the Drone-Action dataset, it surpasses them by a marked margin.

The rest of the study is structured as follows: An overview of human action recognition, including human activity levels and video categories related to HAR, is provided in Section II. The related literature study is illustrated in Section III. A comprehensive study of UAV-based HAR datasets involving their description and classification, UAV challenges, a novel taxonomy derived from in-depth analysis of human activities, and an insightful discussion of the similar and distinct activities within these datasets is introduced in Section IV. The experiment and discussion are displayed in Section V. Finally, the study is concluded in Section VI.

II. Background of Human Action Recognition

Interpreting and understanding human actions are an urgent need and prerequisite for a wide range of reality applications. An awareness of hazardous human activities is required for surveillance system applications [18], monitoring the activity of falls in the elderly is mandatory for health monitoring systems [24], and many applications of such technologies are based on human behavior understanding, such as video retrieval, content-based video summarization, and human–computer interaction. Defining human behavior at an early stage prevents potentially serious injuries through prompt self-correction or immediate manual intervention [25]. The process of manually identifying such actions can be messy, expensive, and prone to mistakes. Consequently, automated action recognition, aka Human Action Recognition (HAR), has gained popularity and attention from the research scientific community.

In the domains of robotics, artificial intelligence, and computer vision, the HAR field has long been a focus. Earlier approaches to understanding human actions relied on HAR datasets recorded by stationary cameras called ground-based HAR datasets [26]. In ground-based HAR datasets, human subjects usually appear large and in the middle of the video frames, and occupy most of the frame. Although these conditions make HAR relatively explicit and facile, this does not accurately simulate real-world scenarios. When human subjects are far from the cameras, the motion scenes typically occur in a small area within the video frames.

In recent years, the research community has presented a new dimension for HAR where incorporating it with the agility, mobility, and high altitude of UAVs. UAVs have the capability to capture various scenes – often with large dimensions – from different angles and altitudes. This provides extensive data that enhances the performance of the action recognition-based model. The transition from ground-based HAR to UAV-based HAR is not simply due to a change in the altitude [27, 28]; instead, it is a context conversion that yields new challenges to the HAR arena, including motion blur and change in subject sizes, among other factors. How to recognize, understand, and interpret human actions by machines is a highly sensitive and essential issue in the general HAR field, particularly in UAV-based HAR; consequently, it is necessary to recognize human behavior accurately and quickly. UAV-based HAR has various real-world applications in surveillance and security, defense and military, sports, and search and rescue. Subsection IV-B provides a detailed discussion of the challenges facing the application of UAVs.

A. Human Activity Levels

Human activity refers to any gesture, action, or behavior performed by a human, individually or collectively, whether intentionally or unintentionally. It involves a wide range of fields and effects, from daily routine tasks and cultural, economic, and political practices to complex industrial processes. These activities can be categorized based on several taxonomies, such as type, purpose, and complexity. Through a literature review [29, 30], we identified five levels of human activities that rely on their complexity execution, namely human-gesture activity level, single-human action level, human-object interaction level, human-human interaction level, and composite human activity level, as shown in Fig. 1.

1) Human–Gesture activity level: A gesture is a meaningful nonverbal communication based on the movement of the human body—often the arms, head, face, and hands—that intends to convey specific emotions or ideas; it typically

occurs over a short period. It is considered the smallest unit of human activity through which more significant and complex activities can be created. Examples include hand movements, facial expressions, and nodding.

2) Single–human (atomic) action level: An atomic action (AO) is a sequence of relevant gestures combined into a single action that a single person executes at a specific moment in time. It refers to behavior that begins and ends with one person without any participation from an object or another person. Examples include walking, running, and swimming, which are all considered single-human actions.

3) Human–Object Interaction (HOI) level: Many activities are a consequence of the interaction between two actors, one of whom is human and the other either a human or an object. An action that occurs between a human and a physical object is termed an HOI interaction. Examples include reading, cooking, eating, and kicking a ball.

4) Human–Human Interaction (HHI) level: Unlike HOIs, which involve understanding how humans and physical objects interact, human-human interactions focus on communication and interaction between two persons. A person's actions toward another determine the type of HHI interaction, such as handshaking, hugging, and fighting.

5) *Composite human activity level:* This refers to a series of sequential or concurrent activities performed by humans that involve a combination of gesture activities, AO actions, and HOI interactions, along with HHI interactions that occur among more than one human and more than one object. In contrast to simple activities such as walking and sitting, composite activities combine these simpler activities that are organized and significant in which this combination forming more complex behaviors or tasks. Examples include group study, conducting meetings, and presentations.

Fig. 1. Human activity levels on HAR.

B. Video-based HAR Category

Through the investigation and analysis of human action recognition video datasets, together with a review of previous literature [5, 8, 31-33], as shown in Fig. 2, we can sequence human action videos into four categories relying on their contents as follows:

1) Multi-action video

- *Multiple Actions by Multiple Actors (MA-MA) over video.* This category includes scenarios of multiple actors performing various activities concurrently within the same scene. It focuses on comprehending difficult scenarios, including several actors participating in diverse activities. It is usually used for crowd behavior analysis and group activity recognition tasks.

- *Multiple Actions by a Single Actor (MA-SA) over video.* This involves scenarios in which a single actor performs multiple activities sequentially or simultaneously throughout time. The primary objective is to recognize the boundaries and types of each action within a continuous video sequence. Such approaches require identifying the beginning and ending times of each action in a video; as such, this category is commonly applied to complex action recognition tasks.

2) Single-Action video

- *Single Action by Multiple Actors (SA-MA) over video.* This category covers scenarios where several actors perform the same activity, often in a synchronized or coordinated manner. It focuses on identifying activities executed by several people working together in cooperation. Typically applied to the analysis of team sports, crowd management in public settings like colleges and seasonality events, and industrial systems' assembly line operations.

- *Single Action by Single Actor (SA-SA) over video.* This scenario, which focuses on a single actor executing a single action within a video, is the most prevalent in human action recognition and has been extensively covered in previous literature. The objective of this category is to classify the action from the video sequence, which forms the core of more complex scenarios.

Fig. 2. Video-based HAR category.

III. Literature Study

With the rapid development of UAVs across various real-world applications, including surveillance systems, crowd management, and search and rescue (SAR), human action recognition in aerial videos has garnered significant attention

in recent years. In contrast to ground-based footage, UAV-based footage typically involves challenges such as changing object and subject sizes, partial or full occlusion, changing angles of view, and background dynamics, which are factors that greatly complicate the consistent categorization of human behaviors and activities. Amid these challenges, researchers have increasingly focused on automated human action recognition, which aims to identify and understand human behaviors, motions, and interactions with the external environment in an aerial context. Some previous studies have sought to perform one-class anomaly detection, whereas others rely on extremely detailed categorization for building a robust automated human activity recognition system. In contrast, others work on inconsistent or ambiguous categorization of the datasets. The next subsections provide how recent UAV-based works categorize diverse actions, from overlapping and inconsistent taxonomy to single-category taxonomy, as these data sources are often relied upon in human behavior categorizations.

A. Inconsistent Category Practices

Some studies have utilized several UAV datasets, including Okutama-Action [21], UAV-Human [22], NEC-Drone [34], UAV-Gesture [35], Drone-Action [8], and UCF-ARG [36] to recognize human activities. These works have categorized human activities in an inconsistent or overlapping manner: some group two different actions under one category, whereas others categorize the same action under two or more categories. In order to handle the Okutama-Action dataset, for instance, Khan et al. [8] and Yadav et al. [9] introduced a multi-label recognition approach focusing on many human activities performed by several persons simultaneously. Although they categorized many fundamental human activities appropriately, such as walking, hugging, and shaking hands, they nevertheless aggregated the “someone pushing a shovel's container” action and the “someone pushing another person” action under the same category termed “pushing/pulling”. Only one of the human action properties is employed in this categorization; the intentionality and level of activity were disregarded in favor of grouping the action according to its kinematics, which resulted in an inconsistency in their categorization.

The NEC-Drone dataset, among others, was leveraged by Xian et al. [10] to analyze human behavior and categorize human actions into individual or group interactions. They considered human activities such as sitting, running, and drinking as individual interactions, while hugging, handshaking, and pushing a person were considered as group interactions. Nonetheless, they separated the walking activity into two independent categories: “walk” as an individual interaction, and “walk toward each other” as a group interaction, which led to overlapping in their categorization. The same issue occurs in the UAV-Human dataset used by Liu et al. [11], where they divided the “walking” action into four categories: “walk”, “walk side by side”, “walk away from someone”, and “walk toward someone”. Several studies [21, 29, 30] have shown that the regular 'walking' action is considered an atomic action that does not interact with any object or human. Consequently, the human action properties and levels of this act are the same and unchanged whether a group walk (e.g., students walking to the schoolyard) or an individual walk (e.g., a person walking to the garden) is considered. The categorization of the action remains unchanged unless the human behavior level changes. For example, when walking with something, such as a crutch or a wheelchair, or when holding the hand of another person, the categorization of these situations can be considered independent of typical “walking”.

Dhiman et al. [12] introduced an aerial human activity recognition framework that utilizes Drone-Action and UCF-ARG datasets. Despite their accurate categorization of many common behaviors across datasets, their categorization exhibits overlaps because they classify the same activity under many categories. For example, the punching action is categorized into the “punching” category on one dataset and the “boxing” category on the other. Similarly, the jogging action is categorized into “jogging”, “jogging side”, and “jogging front/ back”. Azmat et al. [13] provided a UAV-based HAR approach employing three datasets: UAV-Gesture, Drone-Action, and UAV-Human. All human activities involved in these datasets were utilized, except UAV-Human, from which they chose fifteen human actions. Their categorization also suffers from inconsistency and overlapping, since they combine two different actions under the same category, as well as dividing the same action into different categories. For instance, the wave two-hands action with the wave single-hand action are combined under the “waving hands” category, in which each action has different human action properties, where the first is a gesture to send a distress call for help, while the second is a gesture to say hello or goodbye. Furthermore, the Drone-Action and UAV-Gesture datasets include the same action of waving with both hands; however, they separate it into independent categories termed “wave off” and “waving hands”.

B. Extremely Fine-Grained Category Approach

Some research studies have considered detailed categorization of human activities, which focuses overly on the object involved in the action rather than the movement itself. For instance, Hu et al. [14], Uddin et al. [15], Abbas et al. [16], and Jin et al. [17] employed Action-Drone and UAV-Human datasets to understand human activities in aerial contexts. Although their human actions categorization of the datasets was sufficient, it was overly precise in categorizing some actions. For instance, the “hitting” action was split into subcategories such as “hit with a stick” and “hit with a bottle”, despite having the same fundamental behavior, but their categorization relies on which object is utilized. Although this may at first seem accurate, in real-world practice, it may greatly hinder the training of robust AI-based models. By focusing more on the object than the action itself in classifying actions, the model becomes overly detail-oriented about the object, whereas if the model encounters a new scenario, such as “being hit with a hammer”, it would likely classify it incorrectly. Furthermore, changing the viewing angle of the UAV footage causes partial or full occlusions that may block the object's view, and the model may fail to recognize the hitting action if it cannot recognize the same object. Consequently, the model’s ability to generalize is significantly constrained.

C. Broadly (One-class) Category Approach

To avoid issues relating to fine-grained categories, some research studies have adopted a one-class category method on the MDVD dataset [37], whereby rare and anomalous activities are grouped into a single category termed "abnormal", while other activities are categorized as "normal". For instance, Chriki et al. [18] proposed a UAV-based one-class anomaly approach, where they categorize any normal behaviors (e.g., walking and talking) alongside suspicious behaviors (e.g., people loitering and looking inside a car) under the normal category. They base their approach on the idea that suspicious actions are inherent to human nature, which is characterized by curiosity and an interest in discovery. While Mehmood [19], merged suspicious behaviors and anomaly behaviors (e.g., fighting, stealing, and crashing) into one category termed "abnormal". They argued instead that suspicious acts cannot be ignored and should be considered the beginning of abnormal behavior. The issue with their categorization was the inability of the AI-based model to distinguish between the severity or priority of anomalous behaviors; for instance, the scene of "looking at the car suspiciously" has the same priority as the scene of "physical assault". In contrast, although Hamdi et al. [20] also employed a one-class method, they excluded suspicious actions from their categories and concentrated on normal and abnormal behaviors. By concentrating on combining all deviant behaviors into a single category, these approaches circumvent the problem of precisely characterizing abnormal behaviors. Although this method works effectively for uncertain real-world anomaly detection systems, it is unable to distinguish between abnormal actions, which leads to sacrificing the ability to interpret crucial scenes and hinders advanced systems that require responding differently to violent scenes compared to benign anomaly scenes.

D. Theoretical and Methodological Gap

This literature review shows that different frameworks have been utilized for human activity categorization in previous UAV-based HAR studies. Some frameworks fail to consider all properties of human action when categorizing actions, for instance, [8, 9] focused only on the kinematic property, while [10-12] disregard all properties in their categorization, leading to inconsistent and overlapping categorizations that make robust AI-based models confused during the learning phase. In contrast, some others [14-17] overlook human activity levels in their categorizations and depend on highlighting the details of the object involved in the action more than the action itself, displaying extremely fine-grained categorizations, which often causes a failure in the model's ability to generalize for unseen but similar behaviors. Meanwhile, the broad one-category anomaly frameworks [18-20] relied only on human behavioral characteristics, resulting in the loss of fine-grained distinctions between actions and obscuring the critical distinction between violent and harmless suspicious behaviors.

Ultimately, the challenges of these classifications emphasize the demand for a uniform, detailed human behavior categorization in the aerial context. This research aims to fill these gaps by proposing a novel criteria-driven unified taxonomy framework derived from a comprehensive analytical study of human behaviors across ten publicly available UAV datasets. This categorization framework is founded on three theoretical criteria: 1) human action properties, to unify visually distinct and behaviorally similar actions through kinematics, intentionality, and context dependency; 2) human behavioral characteristics, to determine normal from abnormal behaviors through persistence and harmfulness; 3) human activity levels, to demonstrate consistent categorization of actions across different social contexts through define gesture, individual actions, interactions, or collective behaviors. In contrast to previous frameworks that rely on a full/partial single criterion, this triple-criterion framework formalizes categorization by fusing behavioral interpretable integral criteria, correlating human movement theory with aerial surveillance perception, standardizing the distinction of categories of human behaviors, and providing a consistent, generalizable categorization across several UAV datasets. A comparative summary of several categorization frameworks from previous UAV-based HAR studies is presented in Table I. The following sections describe in detail the methodology used to construct this categorization framework to achieve robust automated UAV-based HAR.

TABLE I. A COMPARATIVE SUMMARY OF SEVERAL CATEGORIZATION FRAMEWORKS FROM PREVIOUS UAV-BASED HAR STUDIES

Study	Criteria of Taxonomy			UAV dataset used	Comments
	Human Action Properties	***Human Behavioral Characteristics***	***Human Activity Levels***		
[8, 9]	Only Kinematics	×	×	One	Inconsistent category - considered different actions under the same category
[10, 11]	×	×	×	One	Inconsistent category - considered same action under the several categories
[12]	×	×	×	Two	Inconsistent category - considered same action under the several categories
[13]	×	×	✓	Three	Inconsistent category - considered different actions under the same category
[14-17]	✓	×	×	Two	Fine-grained category - focused on the object more than the action itself
[18-20]	×	✓	×	One	Broadly category - sacrificing the ability to interpret crucial scenes
Our Proposed Taxonomy	✓	✓	✓	**Ten**	**Unified precise consistent category - tradeoff between fine-grained and broad categories.**

IV. UAV-Based HAR Dataset: Description, Analysis, and Preparation

In this section, we introduce a comprehensive study of UAV datasets for human action recognition. First, we provide a brief description of UAV datasets and discuss several perspectives, including human activity level, video category, label scheme, event contents, and environments, in Section IV A. Then, we highlight six challenges found across ten UAV datasets in Section IV B. Finally, we present a novel taxonomy derived from a detailed analytical study of human activities and insightful discussion of both similar and distinct activities in these ten datasets in Section IV C.

A. UAV-Based Datasets Description

Fifteen UAV-based datasets were found in the literature, most of which are available online. The description of these datasets is detailed in the following:

The SAR-UAV [5] dataset contains 2000 images of different human actions captured by a camera drone with a height of 10–40 m at an outdoor place (inside and outside on a campus) during different times of the day. It contains different situations of human actions, such as one person performing a specific action, a group of people performing the same action, or a group of people performing different actions. Each image was annotated with a bounding box and labels. The whole dataset consists of two smaller datasets: a six-class actions dataset and a two-class actions dataset. The six-class actions dataset for general action recognition contained six classes of action: standing, walking, sitting, lying down, handshake, and hand-waving. The two-class actions dataset has two classes of actions, which are: hand-waving and others (combined with other actions as one class). The two-class dataset is for recognizing the hand-waving action, which is a special sign to ask for help in search and rescue (SAR), such as a person walking with hand-waving, sitting with hand-waving, etc.

The Mini-Drone Video Dataset (MDVD) [37] contains 17 human actions across 12 different scenarios with 38 videos of 16-24 seconds that contain varying events in outdoor locations (car parks) captured by a drone camera at different heights and at different times of the day (morning and night). The videos are classified into three groups: normal events (people walking, talking with each other, riding in cars or parking cars), abnormal events (people fighting, stealing things from cars, stealing cars or parking their cars incorrectly), and suspicious events (where no wrongdoing occurs, but unusual or dubious behavior takes place e.g. a person take a photo of the parked cars, people talking surreptitiously). The video frames were annotated with box bounding and action labels.

The Aerial Violent Individual (AVI) [4] dataset contains 2000 images of different human actions captured by a drone camera with a height of 2–8 m in outdoor locations (public places such as parks, streets, and the roof of a house). Each image contains two-ten persons, for which 48% of human actions in the dataset include five aggressive actions such as strangling, stabbing with a knife, punching with a hand, kicking, and shooting with a gun. Each image is associated with a human pose as a labeled annotation.

The VisDrone2018-VID [38] dataset contains 96 videos of ten object classes of interest, such as a person, car, bicycle, and bus, captured by different drone cameras in outdoor locations (streets, parks, walkways, buildings, and bridges in three cities in China) at different times of day. The total number of frames in this dataset is 33400, and is associated with box bounding and the object class label as a labeled annotation.

The Drone-Action [39] dataset contains 240 videos with 66919 frames of 13 human action categories including walking, scampering, running, punching, beating with a bottle, beating with a stick, kicking, stabbing, applauding, and hand waving captured by a drone camera with a height of 8–12 m in the morning in outdoor location (unpaved road in a wheat field). The dataset contains 10 actors, and each action category is repeated five to ten times. Each video was annotated with box bounding, action category labels, and actor ID.

The Okutama-Action [21] dataset contains 43 videos with 77365 frames of 12 human actions captured by a drone camera with a height of 10–45 m in an outdoor location (baseball field) at different times of the day (morning and noon). The dataset divides the 12 actions into three groups: human-human interactions (handshake and hug), human-object interactions (read, push/pull, call, drink, carry), and non-interactions (walk, run, sit, stand, lying down). The dataset contains nine actors, each of whom performs different actions at the same time (e.g., carrying while walking, calling, and handshaking while standing). Two UAV cameras are used to capture the same scenarios at the same time from different angles (top, left, and right) and in different regions of the same location. Each frame contains 0–9 actors, and each video is annotated with box bounding and action labels.

The DroneSURF [40] dataset contains 200 videos with 411451 frames of people walking and talking captured by a drone camera at low height in the morning at outdoor locations (building roofs, parks). The dataset contains 58 actors, and each video contains a group of 2 to 3 actors. Each frame was annotated with a bounding box and annotated face images.

The NEC-Drone [34] dataset contains 5250 videos with more than 460000 frames of 16 human actions captured by two drone cameras at an indoor place (school gym). The dataset contains 19 actors and divides the 16 actions into two groups: one-person interactions (walk, run, jump, sit, talk, drink, throw, carry a backpack) and two-person interactions (handshake, hug, push person, exchange backpack, walk toward each other). Each video was annotated with action labels. Action label annotation was provided for only 2079 videos.

The DroCap [41] dataset contains six videos of three different human activities, including boxing, walking, and playing soccer, captured by a UAV camera at different heights in an indoor location (large room). The dataset contains two actors, and each video contains one actor who performs an action without changing their location. Each video was annotated with the ground truth of the skeleton motion.

The Drone-dataset [42] dataset contains six videos of variant motions with a length of 10 seconds captured by a drone camera in the morning at an outdoor location (parks, groves, and campus). Each video was annotated with a motion object label every five frames. Each frame is associated with the recorded sensing data.

The UAV-Human [22] dataset aims to develop an understanding of human behavior by focusing on four main tasks: action recognition, pose estimation, person re-identification, and attribute recognition. It contains 22476 videos with 119 subjects of 155 human actions for the Action recognition task, 22476 frames for the Pose Estimation task, 41290 frames of 1144 identities for the person Re-Identification task, and 22263 frames for the Attribute Recognition task. This dataset captured videos from 45 different indoor/outdoor locations (farmland, squares, rivers, forests, campuses, gyms, and inside buildings) with different UAV flight attitudes (hover, descent, rotate), varying heights (2–8 meters), various weather conditions (windy, rainy, fog), and at different times of day (morning and night). In addition, it provides multiple data modalities, such as RGB-video, fisheye video, night-vision video, IR sequences, and depth maps. In the Action Recognition task dataset, 155 human actions are clustered into six events: daily events (e.g., wearing a mask), productive events (e.g., fishing), violent events (e.g., stabbing with a knife), social interaction events (e.g., whispering), life-saving events (e.g., calling for help), and Control Gestures (e.g., have command). Each video was annotated with a person ID, gender, clothes, and action labels, and captured timestamps.

The Aerial-Gait [43] dataset contains 17 videos captured by a drone camera with a height of 10–45 m at an outdoor location (park) for the gait recognition task. This dataset has only one human-atomic action, namely walking, which is separately recorded in different situations with two actors: 1) moving drone towards or away from subjects; 2) walking on one circle with different heights; 3) walking in two circles with a fixed height.

The UAV-GESTURE [35] dataset contains 119 videos of 13 human gesture categories as follows: all clear, have command, hover, land, landing direction, move ahead, move downward, move to left, move to right, move upward, not clear, slow down, and wave off. These are captured by a drone camera at an outdoor location (unpaved road in a wheat field) in the morning for the gesture recognition task. The dataset contains 10 actors and each gesture category is repeated five–ten times. Each video was annotated with bounding boxes, body joints, and gesture labels.

The P-DESTRE [28] dataset contains 75 videos with 269 subjects captured by a UAV camera with a height of 5.5–6.7 m at an outdoor location (crowded campus) for detection, tracking, re-identification, and search for pedestrians. This dataset has 14 human action, as follows: walking, running, standing, sitting, cycling, exercising, petting, talking on the phone, leaving a bag, dating, trading, offending, fall, and fighting. Each video was annotated with a person ID and 16 soft biometrics labels (i.e., bounding box, age, wear clothes, and action labels).

The UCF-ARG [36] dataset contains 1442 videos with 12 subjects captured by three cameras (aerial, rooftop, and ground cameras) with a height of 100 feet at an outdoor location (car park on a campus). The dataset consists of two small sets: a ten-class actions set and a seventeen-class actions set. The ten-class action set has 1440 videos with a length of 10 seconds that contain boxing, carrying, clapping, digging, jogging, opening and closing trunk, running, throwing, walking, and waving. The seventeen-class action set has two videos with a length of 1 to 3 minutes that contain the standing, picking-up, gesturing, tennis swing, closing trunk, opening trunk, and jump, including the ten actions of the previous set. Each video was annotated with person ID and action labels.

1) Discussion on the UAV-based HAR datasets: Table II and Table II provide a comprehensive study of the fifteen UAV video datasets found in the literature from many perspectives, as follows:

- Dataset purpose. Datasets focus on various purposes, such as abnormal events detection in [4, 37], human action recognition [5, 21, 22, 34, 36, 39, 41], face recognition [40], object detection [38, 42], gait recognition [43], gesture recognition [35], and detection, tracking, re-identification and search for pedestrians [28].
- Human Activity Levels. For the human activity level, most datasets contain a combination of activity levels, such as the HOI, HHI, and AO levels [21, 28, 34, 37]; or the HOI and AO levels [36, 39]. On the other hand, another study [22] has a composite of four activity levels, while others only contain the AO level [43] or the gesture level [35].
- Data and event content. Most datasets are video datasets [21, 22, 28, 34-43], while others are image datasets [4, 5]. Additionally, some of these datasets contain normal and abnormal events [4, 21, 22, 34, 36, 37, 39], while others contain only normal events [5, 28, 35, 41, 43].
- Label schemes and video categories. Datasets can focus on variant label schemes such as single label per video [22, 34-36, 39, 43] or multiple labels per video [5, 21, 28, 36, 37]. Additionally, some datasets contain MA-MA videos [4, 21, 36, 37], some contain MA-SA videos [34, 37], some contain SA-SA videos [22, 34, 35, 36, 39, 41, 43], and others contain SA-MA videos [5, 28].
- UAV Challenges. Many datasets [4, 5, 21, 22, 28, 34-40, 43] consider challenges such as scale variability, motion blur, UAV perspectives, various illuminations, and full/partial occlusions.
- Environment. All datasets considered were recorded in an outdoor environment, except datasets [34, 41], which were recorded in an indoor environment. Dataset [22] was recorded in both environments. Furthermore, most datasets were recorded in one place [21, 28, 34-37, 39, 41, 43], whereas others were recorded in two to

five different places [4, 5, 38, 40, 42] except [22] was captured in 45 different places.

- Climate conditions and UAV Attributes. All these datasets were recorded under one climate except datasets [21, 22, 42], which were recorded under different weather conditions, including sunny, cloudy, windy, and rainy conditions. Almost all datasets were captured via varied UAV attitudes except [35], which was captured at a fixed attitude. The UAV speed was varied in some datasets [21, 22, 36, 37] and fixed in others [5, 35, 39, 41, 43]. For UAV altitude, all these datasets were recorded under varied altitudes except [41], which was obtained at a fixed altitude.

Furthermore, all studied datasets are available online except for datasets [4, 41, 42]. Datasets [4, 38, 40-42] are excluded because some of them [38, 40] lack any recorded activities or events, whereas others [4, 41, 42] are unavailable online. Consequently, ten of the fifteen UAV-based HAR datasets—SAR-UAV [5], MDVD [37], Drone-Action [39], Okutama-Action [21], NEC-DRONE [34], UAV-Human [22], Aerial-Gait [43], UAV-GESTURE [35], P-DESTRE [28], and UCF-ARG [36]—are available to the research community and contain normal or abnormal events; this study continues using only these datasets.

B. UAV-Based Video Challenges

Through a comprehensive study of the nature of videos in the UAV-based datasets, as shown in Table II and Table III, we found two important compounds related to the circumstances recorded in the video, which form and define the UAV-based challenge types. The first compound is the real-world environment captured on video, such as the number of recorded sites, whether they are indoor or outdoor locations, and the nature of the location observed. This plays a fundamental role in shaping the challenges related to the video backgrounds. The second compound includes factors related to weather conditions and UAV attributes such as attitudes, altitudes, and speed. These components constitute challenges related to resolution and clarity. For instance, a UAV camera's speed influences image resolution, resulting in blurry motion. Different UAV altitudes and UAV attitudes, such as hovering and rotating, impact the subject's size, occlusions, illumination, and the vision's view, as well as weather conditions changing from windy to sunny and cloudy. UAV-based video challenges represent an obstacle to understanding and analyzing videos for different computer vision tasks, such as video summarization and action recognition. After the study had been completed using these two compounds, six UAV challenges were classified in the ten datasets, as shown in Fig. 3 and Fig. 4.

1) Scale variability: UAV cameras capturing videos at vastly different altitudes, leading to various sizes of objects and subjects within the same scene.

2) Low resolution/blurry: These two factors cause loss of information in videos, particularly when using a high altitude of UAV flight or depth of field (DoF) of UAV cameras. When the UAV flies at an altitude above 40 m, subject sizes will appear very small, causing low video resolution. Using a Shallow/Deep DoF on UAV cameras can cause subjects to appear unclear and blurry when the subjects are shown outside the range of the DoF.

3) Motion blur: Sometimes, UAV cameras suffer from being unstable natural cameras; this is primarily caused by the high speed of flight factor, or by climatic factors such as rainfall or strong winds, which lead to motion blur in videos. On the other hand, the sudden and rapid movement of the subjects in the video can also cause motion blur.

4) Fully/Partial occlusions: The UAV camera has a factor distinguished by its wide coverage of the environment from a multi-view aspect, which leads to the appearance of multiple objects along with multiple subjects at the same time in the same scene. This factor sometimes leads to the obscuration or blockage of a subject's body appearing in the scene. This is the most significant challenge facing UAV applications.

5) Various illuminations: Because the conditions of the outdoor environments change (due to changes in weather or climate), subjects in the scenes can be influenced by various lighting, typically containing large shadowed areas of other objects in the same scene, such as buildings, cars, trees.

6) *UAV perspectives:* UAV cameras have features capable of changing the angle of the camera to observe the subject in its current location. However, when the Gimbal pitch indicator of the UAV camera is close to 90 degrees, the long axis of subjects approximately parallel to the UAV camera axis can cause top viewing of subject heads. Consequently, subjects within the scene appear as points for which biometric information is almost non-existent.

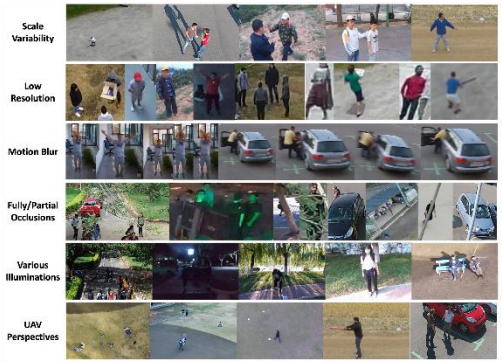


Fig. 3. UAV-based video challenges.

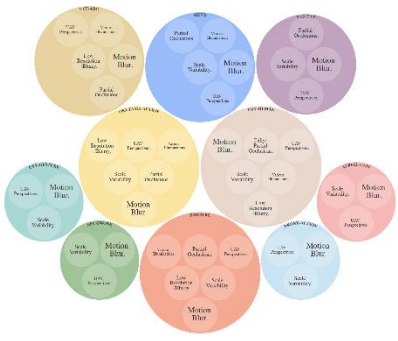


Fig. 4. Challenges on ten UAV datasets.

TABLE II. POPULAR UAV-BASED DATASETS USED IN PREVIOUS STUDIES WITH SPECIFICATIONS AND CLASSIFICATIONS

Dataset/ Year	Actions /Subjects /Videos	Purpose	Human Activity Level	Event Content	Label Scheme	Video Category	Resolution	Environment			Factors				Challenges
												UAV Attributes			
								# sites	Indoor/ Outdoor	Where	Weather conditions	Attitudes	Altitudes	Speed	
SAR-UAV / (2020) [5]	7/-/5	Human Action recognition for Search & Rescue (SAR)	HHI/ AO/Gesture Level.	Normal events	Multi-labels	SA-MA	1920×1080 HD	2	Outdoor	Inside and outside Campus	Windy	Varied	Varied	Fixed	Scale Variability. Motion Blur. UAV Perspectives. Partial Occlusions.
MDVD/ (2015) [37]	17/-/38	Abnormal Event Detection	HHI/ HOI/AO Level.	Both	Multi-labels	MA-MA, MA-SA	1920×1080 HD	1	Outdoor	Car Parking	Sunny	Varied	Varied	Varied	Scale Variability. Motion Blur. UAV Perspectives. Various Illuminations. Partial Occlusions.
AVI/ (2019) [4]	-	Action Aggressive Detection	-	Both	-	MA-MA	1280×720 HD	3	Outdoor	Public Places such as parks, streets, the roofs of a house.	-	Varied	Varied	-	Motion Blur. UAV Perspectives. Various Illuminations. Low Resolution /Blurry.
VisDrone 2018-VID/ (2018) [38]	-	Object Detection	-	-	-	-	3840×2160 HD	5	Outdoor	3 cities in China (streets, parks, walkw	Varied	Varied	Varied	-	Scale Variability. Motion Blur. UAV

										ays, buildings and bridges)					Perspectives. Various Illuminations.
Drone-Action/ (2019) [39]	13/10/240	Human Action Recognition	HOI/AO/Gesture Level.	Both	Single-label	SA-SA	1920×1080 HD	1	Outdoor	Unpaved Road in a Wheat Field	Windy	Varied	Varied	Fixed	UAV Perspectives. Motion Blur. Scale Variability.
Okutama-Action/ (2017) [21]	13/9/431	Human Action Detection	HHI/ HOI/AO Level.	Both	Multi-labels	MA-MA	3840×2160 HD	1	Outdoor	Baseball field	Sunny, Cloudy	Varied	Varied	Varied	Scale Variability. Motion Blur. UAV Perspectives. Various Illuminations. Partial Occlusions. Low Resolution /Blurry.
DroneSUR/ (2019) [40]	-	Face Detection and Recognition	-	-	-	-	1280×720 HD	2	Outdoor	The building roofs, Parks	-	Varied	Varied	-	Scale Variability. Motion Blur. UAV Perspectives. Various Illuminations. Partial Occlusions.

TABLE III. POPULAR UAV-BASED DATASETS USED IN PREVIOUS STUDIES WITH SPECIFICATIONS AND CLASSIFICATIONS (CONT.)

Dataset/ Year	Actions /Subjects /Videos	Purpose	Human Activity Level	Event Content	Label Scheme	Video Category	Resolution	Environment			Factors				Challenges
								# sites	Indoor/ Outdoor	Where	Weather conditions	UAV Attributes			
												Attitudes	Altitudes	Speed	
NEC-DRONE/ (2020) [34]	16/19/2,079	Human Action Detection	HHI/ HOI/AO Level.	Both	Single-label	MA-SA, SA-SA	1920×1080 HD	1	Indoor	School Gym	-	-	Varied	-	Scale Variability. Motion Blur. UAV Perspectives.
DroCap / (2018) [41]	-	Human Motion Capture	-	Normal events	-	SA-SA	-	1	Indoor	Large Room	-	Varied	Fixed	Fixed	-
Drone/ (2017) [42]	-	Moving Object Detection	-	-	-	-	640×480 SD	3	Outdoor	Parks, Groves, and Campus	Sunny, Cloudy	Varied	Varied	-	-
UAV-Human / (2021) [22]	155/119/22,476	Human Behavior Understanding	HHI/ HOI /AO/Gesture Level.	Both	Single-label	SA-SA	1920×1080 HD (RGB)	45	Indoor/ Outdoor	45 places e.g., farmland, squares, rivers	Windy, Rainy	Varied	Varied	Varied	Scale Variability. Motion Blur. UAV Perspectives. Fully/ Partial Occlusions. Various Illuminations. Low Resolution /Blurry.
Aerial-Gait / (2018) [43]	1/2/17	Gait Recognition	AO Level	Normal	Single-label	SA-SA	1920×1080 HD	1	Outdoor	Park	Cloudy	Varied	Varied	Fixed	Scale Variability. Motion Blur. UAV Perspectives.
UAV-GESTURE/ (2018) [35]	13/10/119	Gesture Recognition	Gesture Level	Normal	Single-label	SA-SA	1920×1080 HD	1	Outdoor	Unpaved Road in a Whea	Windy	Fixed	Varied	Fixed	Scale Variability. Motion Blur. UAV

										t Field					Perspectives .
P-DESTRE / (2020) [28]	14/269/75	Pedestrian Detection task, Pedestrian Tracking task, Pedestrian Re-Identification task and Pedestrian Search task	HHI/ HOI/AO Level.	Normal	Multi-labels	SA-MA	3840x2160 HD	1	Outdoor	Campus crowded	-	Varied	Varied	-	Low Resolution/ Blurry. Motion Blur. Partial Occlusions. Various Illuminations. UAV Perspectives. Scale Variability.
UCF-ARG/ (2010) [26]	17/12/473	Action Recognition	HOI /AO/Gesture Level.	Both	Single-labels, Mulit-label	MA-MA, SA-SA	1920× 1080 HD	1	Outdoor	Car parking on the campus	Sunny	Varied	Varied	Varied	Low Resolution/ Blurry. Motion Blur. Various Illuminations. UAV Perspectives. Partial Occlusions.

[1]The earlier description in this section mentioned only 12 activities; subsection IV-C provides additional activity that relies on further analysis.

C. Human Activities Taxonomy on Ten UAV Datasets

In this section, we present a novel unified categorization that relies on a comprehensive study of human activities on ten UAV datasets, which is then divided into three phases. We provide a review and analysis of human activities on UAV datasets, including distinguishing between normal and abnormal behaviors, and determining common and distinct activities in the first phase. The second phase involves grouping the common human activities determined in the first phase under major categories and classifying them according to human activity level. We highlight distinct human activities exclusive to abnormal behavior in particular UAV datasets, in the final phase. The next subsection provides details on the criteria at the foundation of our unified categorization.

1) Criteria-based taxonomy: Our proposed taxonomy is established using three essential criterion sets for human activities from a computer vision perspective: human action properties, human behavioral characteristics, and human activity levels (see Section II A). Each set focuses on a certain aspect of how people perform actions and how these actions are considered normal or abnormal conduct. These criteria are inspired by recent research in human activity recognition and abnormal event detection [44-50], which highlights the importance of movement patterns, activity levels, and situational context in distinguishing one action from another. Together, these criteria allow for systematically unifying or separating human activities throughout the datasets.

a) Human action properties: Numerous studies have demonstrated that the properties of human actions, such as kinematics, intentionality, and context dependency, can be robust indicators of action category rather than a focus on superficial differences, such as the type of object used in the action [45, 46, 50]. Motivated by this, our taxonomy relies primarily on this criterion set to discover similar or distinct actions, when deciding which actions should be combined into a single category, or which should be independent and separate entry.

- Kinematics, the motion pattern or physical standard for human body movement, includes speed, angles, and repetitive cycles. For instance, the actions of “push” and “punch” appear to be almost the same as the arm thrust movement from a UAV's top-down view, i.e.,

they appear to be two similar kinematics. Consequently, they can be unified under a broader category, such as "fighting". Contrarily, if a new action has a completely different kinematic signature, e.g., "hitting with a stick", it remains in a distinct category.

- Intentionality, where the purpose or goal behind the actions varies. For instance, similar motions can have different intents, such as a "hug" versus a "strangle". In UAV footage of both "strangling" and "hugging", a person's arms are wrapped around another person's upper body from behind. However, the intent of the action differs: the former aims to harm, whereas the latter is meant to comfort and show affection. Therefore, these two actions are classified into separate categories due to their distinct intents. In contrast, there are various motions with similar intent, such as a "bow" versus a "handshake". Although the two actions exhibit different body movements, i.e., a slight forward bow compared to extending the arms to clasp hands, they share the same purpose: to greet the person opposite them. Thus, these two actions can be unified under one category, which is "greeting".
- Context Dependency, such that human actions change their meaning depending on the location, the surrounding environment, or the situational context. For instance, a punching action in a wrestling ring would be considered normal in boxing as a popular sport, whereas punching in a public place would be considered violence. This demonstrates that context often radically changes the accuracy of the classification.

b) Human behavioral characteristics: Although the properties of human actions guide the unification of similar motion patterns, the characteristics of behavioral dimensions determine which actions are considered normal or anomalous. Based on experiences from studies on the detection and recognition of abnormal events [47-49], two primary criteria of human behavioral characteristics—persistence and harmfulness—are used to define abnormal human activities.

- Persistence, the duration of human behavior can vary significantly, and is often characterized by repetition and continuity. Normal behavior is defined by its frequency and continuity (e.g., daily routine activities), whereas abnormal behavior is characterized by its rarity, suddenness, and intermittency.
- Harmfulness, the extent to which the act threatens or harms the person or himself. For instance, activities involving aggressive, assaultive, offensive, criminal, or out-of-control conduct, such as falling, are identified as abnormal behaviors, while other behaviors are identified as normal.

2) Phase 1: Human activities review and analysis: Through study and examination of the ten UAV datasets outlined in Section IV-A, we observed that several human activities are similar among multiple datasets or distinct to a specific dataset. As mentioned in the previous section, our taxonomy concluded that similar and distinct human activities are based on commonalities in the properties and levels of human actions. In addition, this approach considers only activities in the context of a public place in the ten UAV datasets, ignoring any activities that have the same action but for a different context, such as a sports context. For example, boxing is a well-known sport with normal behavior in a specific context; however, since the context of these datasets is a public place, boxing activities are considered violent and harmful.

In Appendix A, Table III and Table IV shows an in-depth analysis of the human activities in ten UAV datasets that show normal behaviors, abnormal behaviors of whole activities, and common activities found across multiple datasets, in addition to the video counts and action counts for each action, and the list of actions that occur in each scenario. For example, using the MDVD dataset, the attack scenario contains a person who pushes the driver outside the car and steals the vehicle; this scenario features attacking, running, and stealing activities. The following observations and suggestions were made:

MDVD dataset: after checking the annotation of the videos, we found that some of these actions needed to be modified in the annotation. For example, the crash scenario includes walking, cycling, running, and picking-up actions, but there is no crashing action; we used OpenCV to obtain the start and end frames of this action and add them to the annotated data. According to [18, 51], it is recommended to classify suspicious events as normal events because they are considered part of human nature. In the original dataset, looking inside the car and taking a photo of the car are classified as suspicious; however, it is stated that these behaviors result from human curiosity, exploration, and wandering. Consequently, we consider these suspicious scenarios to in fact be normal behavior. We found that twelve activities in twelve scenarios can be classified as normal behaviors, such as walking, running, picking up, talking, standing, loitering, and bad parking. Furthermore, five activities in seven scenarios can be classified as abnormal behaviors, such as attacking, stealing, falling, fighting, and crashing.

Drone-Action dataset: we identified eight activities that can be classified as normal behaviors, such as clapping, jogging, running, walking, and hand waving. Furthermore, five activities can be classified as abnormal behaviors, such as boxing, hitting with a bottle, hitting with a stick, kicking, and stabbing.

Okutama-Action dataset: after thoroughly investigating human actions, we identified a new action that can be classified under the HHI level. All pushing/pulling actions in the dataset were classified under the HOI level. Upon reviewing the annotations in tandem with their corresponding videos, we found that the two types of activities on this action can be classified based on different human activity levels and human behavior properties. The first shows a person pushing and pulling with a shovel's container, which is normal behavior and can be classified under the HOI level, as mentioned in previous studies. The second action was discovered shows two persons pushing each other, which is

abnormal behavior that can be classified under the HHI level, as shown in Fig. 5. Utilizing the OpenCV package, we identified the start and end frames of this action and integrated them into the annotations accordingly. Therefore, currently, there are thirteen human actions in this dataset instead of twelve. We classified twelve activities as normal behaviors: handshaking, hugging, reading, drinking, pushing/pulling (HOI), carrying, calling, running, walking, lying, sitting, and standing. One activity can be classified as abnormal behavior, namely pushing/pulling (HHI).

Fig. 5. Samples of HHI pushing/pulling on the Okutama-Action dataset.

NEC-Drone dataset: we identified fifteen activities that can be classified as normal behaviors, such as walking, running, jumping, picking up a backpack, and going and leaving a backpack and going, among others. Furthermore, one activity can be classified as abnormal behavior, namely, pushing a person.

UAV Human dataset: we identified 136 activities that can be classified as normal behaviors, such as drinking, eating snacks, writing, and applauding, among others. Furthermore, nineteen activities can be classified as abnormal behavior, such as punching someone, pushing someone, and stealing something from another's pocket. We observed that some activities did not simulate the real-world situation, such as the "rob something from someone" action, which is more akin to taking something in a friendly manner rather than robbing, and the "kick something" action, which is more akin to kicking nothing or kicking a box. The "take a phone for someone" action is similar to taking a photo for someone, and the "slap someone on the back" action is similar to warning someone about something they dropped, although it should be considered more of a warning than an aggressive strike, which can probably be normal behavior. The "chase someone" action appears as a friendly pursuit, which is similar to playing rather than a real chase that suggests danger or harm to the other person.

Aerial Gait dataset, one action, walking, can be classified as normal behavior.

UAV-GESTURE dataset, the whole activities of the Gesture Signal were classified as normal behaviors, such as all clear, hover, and move ahead, among others.

P-DESTRE dataset, after examining the videos, only four out of fourteen actions have a corresponding video. These four activities are classified as normal behavior: walking, standing, sitting, and talking over the phone.

UCF-ARG dataset, we identified sixteen activities that can be classified as normal behaviors, such as throwing, standing, and walking, among others. Furthermore, one activity can be classified as abnormal behavior, namely boxing. We observe that actions are considered to be human-atomic actions (AOs), where the person throws nothing, meaning there is no object in their hands.

SAR-UAV dataset, we identified 122 photos of the Okutama-Action dataset out of 2000 photos in this dataset. To avoid duplicating the samples in this study, we removed these photos from the dataset and retained the 1880 photos created by the authors. We identified five activities that were classified as normal behaviors: standing, walking, running, handshake, and hand-waving. Data are provided as individual frames rather than sequences of frames. Therefore, we aimed to cluster similar frames to convert them into consistent videos. To achieve this, we utilized the pre-trained VGG16 model to extract deep features from the frames. Subsequently, we calculated the cosine similarity among these extracted features. Finally, the frames were clustered into five distinct groups based on their cosine similarity using an unsupervised learning technique called the agglomerative clustering algorithm.

3) Phase 2: Unification of common human activities: Following the first phase, 147 similar human activities out of 266 human activities were identified across the ten datasets. Among these 147 activities, we classified 31 as abnormal behavior based on their nature, and the remaining 116 as routine activities. The unification of the common human activities associated with their human activity level under the precise categories for each dataset is provided in Appendix A [see Table V and Table VI]. Fig. 6 reflects Tables VII and VIII, which highlight the human activity level for each action. The unified, precise categories of the common human activities found were introduced as follows:

a) Abnormal in the unified precise category

- Fighting depicts scenarios involving a physical confrontation, such as scuffles and affrays, including dragging, pushing, punching, and kicking. Sixteen similar activities were identified in the MDVD, Drone-Action, Okutama-Action, NEC-Drone, UAV-Human, and UCF-ARG datasets. This category falls under the AO, HHI, and Composite (HHI+HOI) levels, where these activities have the same human action properties with different human activity levels.
- Robbery depicts scenarios of theft and pickpocketing. Four similar activities were identified in the MDVD and UAV-Human datasets. This category falls under the HOI and Composite (HHI+HOI) levels, where these activities have the same human action properties with different human activity levels.
- Vehicle Theft depicts scenarios such as stealing a car by attempting to pick a locked door or opening it aggressively or stealthily. This category falls under the HOI level. Three similar activities were identified in the MDVD and UAV-Human datasets.
- Assault depicts scenarios involving physical violence with sharp objects, perhaps under threat, such as hitting

with a bottle, stabbing with a knife, and threatening with a gun. Eight similar activities were identified in the Drone-Action and UAV-Human datasets. This category falls under the HOI and Composite (HHI+HOI) levels, where these activities have the same human action properties with different human activity levels.

b) Normal in the unified precise category

- Walking is a form of body movement that includes alternating movement between the legs, including movement forward, backward, and sideways. This category falls under the AO level. Twenty-one similar activities were identified in nine datasets, with the exception of the UAV-GESTURE dataset.
- Clapping is a simple physical activity that results from hitting the palms of the hands together at different rhythms. This category falls under the AO level. Three similar activities were identified in the Drone-Action, UAV-Human, and UCF-ARG datasets.
- Digging is an activity that utilizes machines or equipment to remove soil to create a hole. This category falls under the HOI level. Two similar activities were identified in the UAV-Human and UCF-ARG datasets.
- Drinking is an important biological activity in which fluids in objects are consumed through the mouth. This category falls under the HOI level. Four similar activities were identified in the Okutama-Action, NEC-Drone, and UAV-Human datasets.
- Sitting is an activity for body rest, where the buttocks and thighs are placed on a chair, the floor, or a surface. This category falls under the HOI level. Four similar activities were identified in the Okutama-Action, NEC-Drone, UAV-Human, and UCF-ARG datasets.
- Handshaking is an act performed by two people to greet each other or agree on something. This category falls under the HHI level. Four similar activities were identified in the Okutama-Action, NEC-Drone, UAV-Human, and SAR-Drone datasets.
- Standing is the act of keeping the body in a straight position on the feet, relying on the legs. This category falls under the AO level. Nine similar activities were identified in the MDVD, Okutama-Action, NEC-Drone, UAV-Human, P-DESTRE, UCF-ARG, and SAR-Drone datasets.
- Reading is a process that depends on the human eye to understand what is written, whether it is on a piece of paper, a book, a tablet, or even a mobile phone. This category falls under the HOI level. Two similar activities were identified in the Okutama-Action and UAV-Human datasets.
- Running is an aerobic activity in which the body is pushed forward by the feet in a continuous movement and at different speeds. This category falls under the AO level. Eleven similar activities were identified in the MDVD, Drone-Action, Okutama-Action, NEC-Drone, UAV-Human, UCF-ARG, and SAR-Drone datasets.
- Jumping is a movement in which the body is pushed off the ground using the power of the legs and feet. This category falls under the AO level. Four similar activities were identified in the NEC-Drone, UAV-Human, and UCF-ARG datasets.
- Talking is a spoken communication act carried out between two or more persons, either directly or indirectly, sometimes using tools such as a mobile phone. Eight similar activities were identified in the MDVD, Okutama-Action, NEC-Drone, UAV-Human, and P-DESTRE datasets. This category falls under the HHI and HOI levels, where these activities have the same human action properties with different human activity levels.
- Throwing is a physical activity in which a body of water is propelled through the air by the power of movement of the arm, wrist, and hand. Six similar activities were identified in the NEC-Drone, UAV-Human, and UCF-ARG datasets. This category falls under the AO and HOI levels, where these activities have the same human action properties with different human activity levels.
- Carrying is a physical effort in which an object is transported from one place to another by human body parts, including the hands, shoulders, and head. This category falls under the HOI level. Three similar activities were identified in the Okutama-Action, UAV-Human, and UCF-ARG datasets.
- Hugging is an activity in which two or more people cuddle each other by wrapping their arms around each other. This category falls under the HHI level. Three similar activities were identified in the Okutama-Action, NEC-Drone, and UAV Human datasets.
- Pick-Up is a physical activity in which something is taken, collected, or lifted using the hands. This category falls under the HOI level. Six similar activities were identified in the MDVD, NEC-Drone, UAV Human, and UCF-ARG datasets.
- Exchange Something is an act between two persons in which an object is given and another is received in return. This category falls under the Composite (HHI+HOI) level. Three similar activities were identified in the NEC-Drone and UAV Human datasets.
- Gesture Signals are a form of non-verbal communication in which human body parts (e.g., the arm) are used to inform a specific signal or command, such as all clear, have command, land, hover, etc. This category falls under the Gesture level. Sixteen similar activities were identified in the Drone-Action and UAV Human, UAV-GESTURE, UCF-ARG, and SAR-Drone datasets.

4) *Phase 3: Distinct abnormal human activities:* Following the first phase, five distinct human activities were determined on the MDVD and UAV-Human datasets. Distinct abnormal human activities and their associated human activity levels are provided in Appendix A, Table IX. They are as follows:

- Stagger and Fall are similar activities, where the stagger is the beginning of the fall event. This is a part of the fall event, but not the entire event; for this reason, we decided not to include it in the same category.
- Crashing and bumping into someone are similar activities, but both have only one intentionality, i.e., two people walk in opposite directions and eventually collide. However, they differ in human activity levels and other human activity properties, such as kinematics and context dependency. The first activity involves a person riding a bicycle and another person walking, whereas the second activity involves two people walking. As a result, they cannot be classified in the same category.
- Chest Discomfort is a reaction in which a person raises their hand to the chest area in response to a sensation of pain, tightness, or discomfort. It is classified as an abnormal behavior since it is perhaps dangerous and requires urgent medical intervention.

V. Experiment and Discussion

In this section, we conduct a comparative experiment utilizing a baseline HAR model on both the proposed unified precise taxonomy and other existing taxonomies applied to the benchmark Drone-Action dataset, representing a quantitative evaluation and validation.

A. Experimental Setup

In this experiment, we used Drone-Action dataset, one of the ten UAV datasets employed within the proposed unified taxonomy framework. The ImageNet pretrained MViTv2 model [52], a famous vision transformer used for action recognition and video understanding tasks, was trained on two categorization frameworks under the same training settings. The categorization frameworks involved on the experiment include the original fine-grained categorization found in the literature [14-17] and the proposed unified precise categorization framework, as described in Table IV. The MViTv2 model was trained for 22 epochs using a batch size of 2 and a learning rate of 1e-4, with a split ratio of 70% for the training set and 30% for the testing set. The input videos of the dataset were resized to 224×224 resolutions. The experiment was executed using PyTorch 2.5 with CUDA 12.4 on a device provided with an AMD EPYC 7402 CPU, 24 cores / 48 threads, 2.8 GHz base, a NVIDIA GeForce RTX 3090 GPU, and 24.4 GB HBM2. The performance evaluation metrics used included accuracy, precision, recall, and F1-score measures.

TABLE IV. The Categorization Frameworks Specification

Dataset	# Videos	Categorization Framework	# Categories	Action Categories List
Action-Drone	240	Original Fine-Grained Taxonomy [14-17]	13	Boxing, Clapping, Hitting with bottle, Hitting with stick, Jogging front back, Jogging side, Kicking, Running front back, Running side, Stabbing, Walking front back, Walking side, and Waving hands.
		Proposed Unified Precise Taxonomy (our)	6	Fighting, Assault, Walking, Clapping, Running, and Hand Waving

B. Quantitative Evaluation and Validation

We compare the proposed unified precise taxonomy with the existing taxonomy (fine-grained category) on the Drone-Action dataset, using the same HAR model and training settings to ensure a fair comparison. The results are displayed in Table V. The fine-grained category framework relies on the object's type (e.g., bottle, stick, knife) or the drone's angle (e.g., side view, back-forward view). Despite the categories' similar action properties and activity levels, separating them into individual categories regarding different object types or angles of view hinders the HAR model training. Consequently, as shown in Table V, the HAR model with a fine-grained category framework achieves 78% accuracy and F1-score, significantly constraining the model's ability to generalize. In comparison, the proposed unified precise taxonomy increases recognition accurate by 18.8% and enhances the F-1 score by 19.4%, resulting in improved HAR model ability to understand motion patterns. This may be attributed to the proposed taxonomy relies on three criteria of human action theory and behavioral, resulting in a consistent and balanced category and decreasing confusion among similar human actions. This allows the model to learn the features of human activity patterns, regardless of the shooting angle, object type, or surrounding environment, making it effective and stable for recognizing unseen similar actions.

TABLE V. Comparison of Original Fine-Grained and Proposed Unified Precise Taxonomies on the Drone-Action Dataset

Categorization Framework	Metrics			
	Accuracy	*Precision*	*Recall*	*F1-score*
Original Fine-Grained Taxonomy [14-17]	0.7859	0.8109	0.7859	0.7794
Proposed Unified Precise Taxonomy (our)	**0.9739**	**0.9742**	**0.9739**	**0.9740**

VI. Conclusion

This research has proposed a novel unified yet precise systematic taxonomy for UAV-based HAR derived from a comprehensive study of ten publicly available aerial datasets, which addresses the issues of fine-grained and inconsistent categorization across multiple UAV-based datasets. To the

best of our knowledge, this study is the first of its kind, providing a detailed and comprehensive analysis of human actions present in UAV datasets. The following advancements were made: 1) we explored the available UAV datasets and provided a detailed comparison based on the content of human actions present in their videos; 2) we identified the challenges of UAV datasets encountered in HAR; and 3) we built a novel category for video-based human action recognition. By basing the proposed taxonomy on three criteria relating to human actions and behaviors, we show how object-centric categories can be unified into precise and broader categories oriented toward behavior, intentionality, and levels, while preserving explicitly distinct anomalous actions for surveillance security-focused recognition.

The proposed taxonomy consists of a three-phase framework that: 1) starts by reviewing and analyzing actions across UAV-based datasets, 2) standardizes activities based on similar characteristics across the three criteria, and 3) separates a set of distinct anomalous actions to ensure that critical events are not overlooked. The unified taxonomy includes 152 human activities, providing a compromise that captures basic movement patterns while simultaneously determining critical threats (both normal and abnormal). For evaluation and validation of the performance, we compare the proposed taxonomy with the existing taxonomy on the Drone-Action dataset. The experiment results reveal that the proposed taxonomy outperforms others by a margin of 19.4% in F1-score, indicating that the proposed taxonomy provides balanced, consistent, and generalized categorization. Despite the proposed taxonomy having several benefits, it faced some limitations: 1) It ignores the distinction between the object types involved in human actions in their categorization, which reduces the accuracy of identifying evidence details in forensic and criminal applications. 2) The current evaluation and validation are limited to one dataset and model. In future work, we plan to expand the evaluation to include many datasets and several famous HAR models.

In the future, we envision this taxonomy as fundamental to more consistent dataset organization and real-time UAV surveillance applications. By reducing the fragmentation of categorization, researchers can train deeper or more flexible models on the amassed data, improving the performance of human action recognition and detection in aerial footage. Ultimately, we believe the proposed taxonomy both simplifies the overall classification and enhances the potential to provide robust and interpretable solutions for drone-based human action recognition.

REFERENCES

[1] P. Jin, L. Mou, G.-S. Xia, and X. X. Zhu, "Anomaly Detection in Aerial Videos With Transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2022, doi: 10.1109/TGRS.2022.3198130.

[2] T. M. Tran, T. N. Vu, T. V. Nguyen, and K. Nguyen, "UIT-ADrone: A Novel Drone Dataset for Traffic Anomaly Detection," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 5590–5601, 2023, doi: 10.1109/JSTARS.2023.3285905.

[3] T. M. Tran, D. C. Bui, T. V. Nguyen, and K. Nguyen, "Transformer-Based Spatio-Temporal Unsupervised Traffic Anomaly Detection in Aerial Videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 9, pp. 8292–8309, Sept. 2024, doi: 10.1109/TCSVT.2024.3376399.

[4] A. Singh, D. Patil, and S. N. Omkar, "Eye in the Sky: Real-Time Drone Surveillance System (DSS) for Violent Individuals Identification Using ScatterNet Hybrid Deep Learning Network," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT: IEEE, June 2018, pp. 1710–17108. doi: 10.1109/CVPRW.2018.00214.

[5] B. Mishra, D. Garg, P. Narang, and V. Mishra, "Drone-surveillance for search and rescue in natural disaster," *Comput. Commun.*, vol. 156, pp. 1–10, Apr. 2020, doi: 10.1016/j.comcom.2020.03.012.

[6] A. Alhothali, A. Balabid, R. Alharthi, B. Alzahrani, R. Alotaibi, and A. Barnawi, "Anomalous event detection and localization in dense crowd scenes," *Multimed. Tools Appl.*, vol. 82, no. 10, pp. 15673–15694, Apr. 2023, doi: 10.1007/s11042-022-13967-w.

[7] J. Liu, X. Liu, M. Qu, and T. Lyu, "EITNet: An IoT-enhanced framework for real-time basketball action recognition," *Alex. Eng. J.*, vol. 110, pp. 567–578, Jan. 2025, doi: 10.1016/j.aej.2024.09.046.

[8] M. Khan, J. Ahmad, A. El Saddik, W. Gueaieb, G. De Masi, and F. Karray, "Drone-HAT: Hybrid Attention Transformer for Complex Action Recognition in Drone Surveillance Videos," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA: IEEE, June 2024, pp. 4713–4722. doi: 10.1109/CVPRW63382.2024.00474.

[9] S. K. Yadav et al., "DroneAttention: Sparse weighted temporal attention for drone-camera based activity recognition," Neural Netw., vol. 159, pp. 57–69, Feb. 2023, doi: 10.1016/j.neunet.2022.12.005.

[10] R. Xian, X. Wang, and D. Manocha, "MITFAS: Mutual Infomation based Temporal Feature Alignment and Sampling for Aerial Video Action Recognition," in 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Jan. 2024, pp. 6611–6620. doi: 10.1109/WACV57701.2024.00649.

[11] J. Liu, B. Yin, J. Lin, J. Wen, Y. Li, and M. Liu, "HDBN: A Novel Hybrid Dual-Branch Network for Robust Skeleton-Based Action Recognition," in 2024 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), July 2024, pp. 1–6. doi: 10.1109/ICMEW63481.2024.10645450.

[12] C. Dhiman, A. Varshney, and V. Vyapak, "AP-TransNet: a polarized transformer based aerial human action recognition framework," Mach. Vis. Appl., vol. 35, no. 3, p. 52, May 2024, doi: 10.1007/s00138-024-01535-1.

[13] U. Azmat et al., "Aerial Insights: Deep Learning-Based Human Action Recognition in Drone Imagery," IEEE Access, vol. 11, pp. 83946–83961, 2023, doi: 10.1109/ACCESS.2023.3302353.

[14] Z. Hu, Z. Pan, Q. Wang, L. Yu, and S. Fei, "Forward-reverse adaptive graph convolutional networks for skeleton-based action recognition," Neurocomputing, vol. 492, pp. 624–636, July 2022, doi: 10.1016/j.neucom.2021.12.054.

[15] S. Uddin, T. Nawaz, J. Ferryman, N. Rashid, Md. Asaduzzaman, and R. Nawaz, "Skeletal Keypoint-Based Transformer Model for Human Action Recognition in Aerial Videos," IEEE Access, vol. 12, pp. 11095–11103, 2024, doi: 10.1109/ACCESS.2024.3354389.

[16] Y. Abbas and A. Jalal, "Drone-Based Human Action Recognition for Surveillance: A Multi-Feature Approach," in 2024 International Conference on Engineering & Computing Technologies (ICECT), Islamabad, Pakistan: IEEE, May 2024, pp. 1–6. doi: 10.1109/ICECT61618.2024.10581378.

[17] P. Jin, L. Mou, Y. Hua, G.-S. Xia, and X. X. Zhu, "FuTH-Net: Fusing Temporal Relations and Holistic Features for Aerial Video Classification," IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–13, 2022, doi: 10.1109/TGRS.2022.3150917.

[18] A. Chriki, H. Touati, H. Snoussi, and F. Kamoun, "Deep learning and handcrafted features for one-class anomaly detection in UAV video," Multimed. Tools Appl., vol. 80, no. 2, pp. 2599–2620, Jan. 2021, doi: 10.1007/s11042-020-09774-w.

[19] A. Mehmood, "LightAnomalyNet: A Lightweight Framework for Efficient Abnormal Behavior Detection," Sensors, vol. 21, no. 24, Art. no. 24, Jan. 2021, doi: 10.3390/s21248501.

[20] S. Hamdi, S. Bouindour, H. Snoussi, T. Wang, and M. Abid, "End-to-End Deep One-Class Learning for Anomaly Detection in UAV Video

Stream," J. Imaging, vol. 7, no. 5, Art. no. 5, May 2021, doi: 10.3390/jimaging7050090.

[21] M. Barekatain et al., "Okutama-Action: An Aerial View Video Dataset for Concurrent Human Action Detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA: IEEE, July 2017, pp. 2153–2160. doi: 10.1109/CVPRW.2017.267.

[22] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "UAV-Human: A Large Benchmark for Human Behavior Understanding with Unmanned Aerial Vehicles," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA: IEEE, June 2021, pp. 16261–16270. doi: 10.1109/CVPR46437.2021.01600.

[23] Y. Yin, M. Liu, R. Yang, Y. Liu, and Z. Tu, "Dark-DSAR: Lightweight one-step pipeline for action recognition in dark videos," Neural Netw., vol. 179, p. 106622, Nov. 2024, doi: 10.1016/j.neunet.2024.106622.

[24] A. Alsadoon, G. Al-Naymat, and O. D. Jerew, "An architectural framework of elderly healthcare monitoring and tracking through wearable sensor technologies," Multimed. Tools Appl., vol. 83, no. 26, pp. 67825–67870, Aug. 2024, doi: 10.1007/s11042-024-18177-0.

[25] P. Pareek and A. Thakkar, "A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications," Artif. Intell. Rev., vol. 54, no. 3, pp. 2259–2322, Mar. 2021, doi: 10.1007/s10462-020-09904-8.

[26] P. Kumar, S. Chauhan, and L. K. Awasthi, "Human Activity Recognition (HAR) Using Deep Learning: Review, Methodologies, Progress and Future Research Directions," Arch. Comput. Methods Eng., vol. 31, no. 1, pp. 179–219, Jan. 2024, doi: 10.1007/s11831-023-09986-x.

[27] S. Kapoor, A. Sharma, and A. Verma, "Diving deep into human action recognition in aerial videos: A survey," J. Vis. Commun. Image Represent., vol. 104, p. 104298, Oct. 2024, doi: 10.1016/j.jvcir.2024.104298.

[28] S. V. A. Kumar, E. Yaghoubi, A. Das, B. S. Harish, and H. Proença, "The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking, and Short/Long-Term Re-Identification from Aerial Devices," IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 1696–1708, 2021, doi: 10.1109/TIFS.2020.3040881.

[29] P. Khaire and P. Kumar, "Deep learning and RGB-D based human action, human–human and human–object interaction recognition: A survey," J. Vis. Commun. Image Represent., vol. 86, p. 103531, July 2022, doi: 10.1016/j.jvcir.2022.103531.

[30] A. Ray, M. H. Kolekar, R. Balasubramanian, and A. Hafiane, "Transfer Learning Enhanced Vision-based Human Activity Recognition: A Decade-long Analysis," Int. J. Inf. Manag. Data Insights, vol. 3, no. 1, p. 100142, Apr. 2023, doi: 10.1016/j.jjimei.2022.100142.

[31] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "A combined multiple action recognition and summarization for surveillance video sequences," Appl. Intell., vol. 51, no. 2, pp. 690–712, Feb. 2021, doi: 10.1007/s10489-020-01823-z.

[32] C. Zhao, D. Du, A. Hoogs, and C. Funk, "Open Set Action Recognition via Multi-Label Evidential Learning," in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2023, pp. 22982–22991. doi: 10.1109/CVPR52729.2023.02201.

[33] D. Kothandaraman, T. Guan, X. Wang, S. Hu, M. Lin, and D. Manocha, "FAR: Fourier Aerial Video Recognition," in Computer Vision – ECCV 2022, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., Cham: Springer Nature Switzerland, 2022, pp. 657–676. doi: 10.1007/978-3-031-19836-6_37.

[34] J. Choi, G. Sharma, M. Chandraker, and J.-B. Huang, "Unsupervised and Semi-Supervised Domain Adaptation for Action Recognition from Drones," in 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA: IEEE, Mar. 2020, pp. 1706–1715. doi: 10.1109/WACV45572.2020.9093511.

[35] A. G. Perera, Y. W. Law, and J. Chahl, "UAV-GESTURE: A Dataset for UAV Control and Gesture Recognition," in Computer Vision – ECCV 2018 Workshops, L. Leal-Taixé and S. Roth, Eds., Cham: Springer International Publishing, 2019, pp. 117–128. doi: 10.1007/978-3-030-11012-3_9.

[36] A. Nagendran, D. Harper, and M. Shah, "UCF-ARG dataset," Center for Research in Computer Vision (CRCV) at the University of Central Florida. Accessed: Dec. 22, 2024. [Online]. Available: https://www.crcv.ucf.edu/data/UCF-ARG.php

[37] M. Bonetto, P. Korshunov, G. Ramponi, and T. Ebrahimi, "Privacy in mini-drone based video surveillance," in 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), May 2015, pp. 1–6. doi: 10.1109/FG.2015.7285023.

[38] P. Zhu et al., "VisDrone-VDT2018: The Vision Meets Drone Video Detection and Tracking Challenge Results," in Computer Vision – ECCV 2018 Workshops, vol. 11133, L. Leal-Taixé and S. Roth, Eds., in Lecture Notes in Computer Science, vol. 11133. , Cham: Springer International Publishing, 2019, pp. 496–518. doi: 10.1007/978-3-030-11021-5_29.

[39] A. G. Perera, Y. W. Law, and J. Chahl, "Drone-Action: An Outdoor Recorded Drone Video Dataset for Action Recognition," Drones, vol. 3, no. 4, p. 82, Nov. 2019, doi: 10.3390/drones3040082.

[40] I. Kalra, M. Singh, S. Nagpal, R. Singh, M. Vatsa, and P. B. Sujit, "DroneSURF: Benchmark Dataset for Drone-based Face Recognition," in 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), Lille, France: IEEE, May 2019, pp. 1–7. doi: 10.1109/FG.2019.8756593.

[41] X. Zhou, S. Liu, G. Pavlakos, V. Kumar, and K. Daniilidis, "Human Motion Capture Using a Drone," in 2018 IEEE International Conference on Robotics and Automation (ICRA), May 2018, pp. 2027–2033. doi: 10.1109/ICRA.2018.8462830.

[42] C. Huang, P. Chen, X. Yang, and K.-T. T. Cheng, "REDBEE: A visual-inertial drone system for real-time moving object detection," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC: IEEE, Sept. 2017, pp. 1725–1731. doi: 10.1109/IROS.2017.8205985.

[43] A. G. Perera, Y. W. Law, and J. Chahl, "Human Pose and Path Estimation from Aerial Video Using Dynamic Classifier Selection," Cogn. Comput., vol. 10, no. 6, pp. 1019–1041, Dec. 2018, doi: 10.1007/s12559-018-9577-6.

[44] A. Ben Mabrouk and E. Zagrouba, "Abnormal behavior recognition for intelligent video surveillance systems: A review," Expert Syst. Appl., vol. 91, pp. 480–491, Jan. 2018, doi: 10.1016/j.eswa.2017.09.029.

[45] W. Zhang et al., "Putting human behavior predictability in context," EPJ Data Sci., vol. 10, no. 1, Art. no. 1, Dec. 2021, doi: 10.1140/epjds/s13688-021-00299-2.

[46] Z. Kabulska and A. Lingnau, "The cognitive structure underlying the organization of observed actions," Behav. Res. Methods, vol. 55, no. 4, pp. 1890–1906, June 2023, doi: 10.3758/s13428-022-01894-5.

[47] L. Zanella, B. Liberatori, W. Menapace, F. Poiesi, Y. Wang, and E. Ricci, "Delving into CLIP latent space for Video Anomaly Recognition," Comput. Vis. Image Underst., vol. 249, p. 104163, Dec. 2024, doi: 10.1016/j.cviu.2024.104163.

[48] R. Zhao et al., "A Review of Abnormal Crowd Behavior Recognition Technology Based on Computer Vision," Appl. Sci., vol. 14, no. 21, p. 9758, Oct. 2024, doi: 10.3390/app14219758.

[49] R. Mojarad, A. Chibani, F. Attal, G. Khodabandelou, and Y. Amirat, "A hybrid and context-aware framework for normal and abnormal human behavior recognition," Soft Comput., vol. 28, no. 6, pp. 4821–4845, Mar. 2024, doi: 10.1007/s00500-023-09188-4.

[50] L. C. Vinton, C. Preston, S. de la Rosa, G. Mackie, S. P. Tipper, and N. E. Barraclough, "Four fundamental dimensions underlie the perception of human actions," Atten. Percept. Psychophys., vol. 86, no. 2, pp. 536–558, Feb. 2024, doi: 10.3758/s13414-023-02709-1.

[51] J. Henrio and T. Nakashima, "Anomaly Detection in Videos Recorded by Drones in a Surveillance Context," in 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Oct. 2018, pp. 2503–2508. doi: 10.1109/SMC.2018.00429.

[52] Y. Li et al., "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 4794–4804. doi: 10.1109/CVPR52688.2022.00476.

APPENDIX A

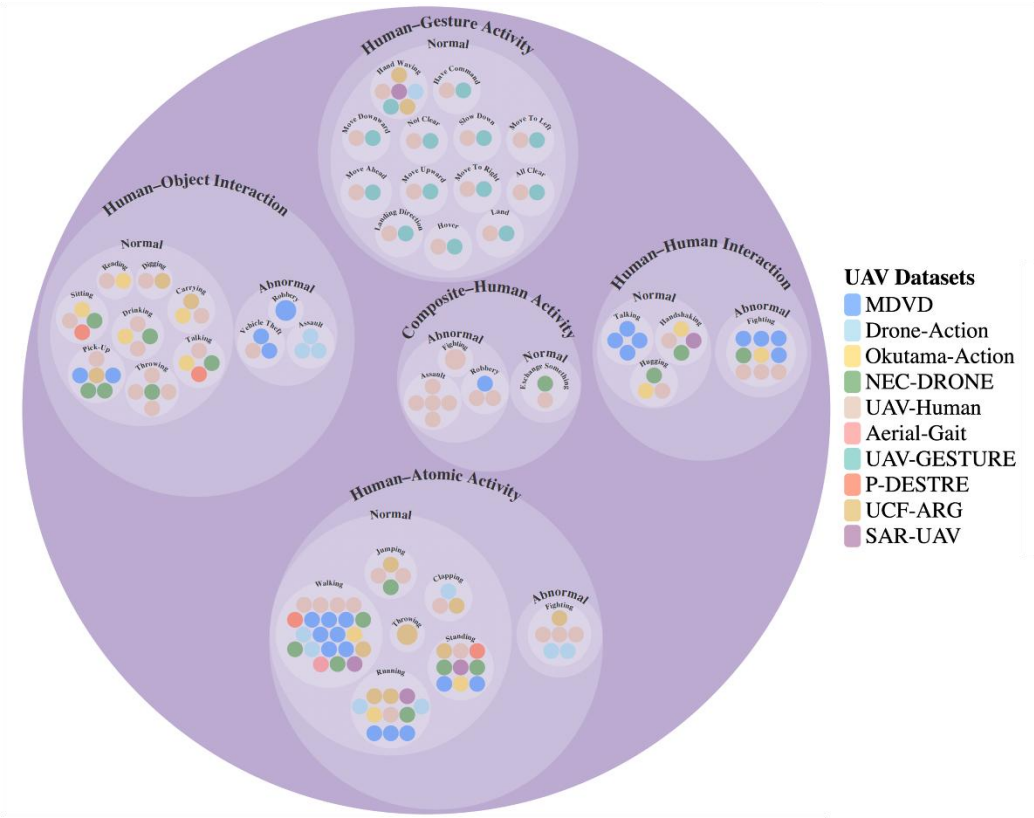


Fig. 6. Unified precise categorization of common human activities on ten UAV datasets.

TABLE VI. HUMAN ACTIVITIES ANALYSIS OF THE UAV DATASETS

UAV Dataset	No. of actions	No. of videos	Normal Behavior	Abnormal Behavior	Common Activities
MDVD	12 scenarios (17 actions)	38	badparking(4; badparking, walking), broken(2; broken),normal(10; walking, talking, standing, normal), reserving(2; walking, loitering, reserving), suspicious(6; suspicious, loitering, talking, walking), attack(1; running), falling(1; walking), fighting(1; talking), crash(2, walking, cycling, running, picking_up), stealingcar(3; loitering), stealinginside(1; running, walking), stealingpedestrian(5; walking, standing, talking, loitering, picking_up)	attack (1; attacking, stealing), falling(1; falling), fighting(1; fighting), crash(2, crash), stealingcar(3; stealing),stealinginside(1;stealing),stealingpedestrian (5;stealing, fighting, attacking)	fighting (talking, fighting), attack (attacking, running, stealing), normal (walking, standing, talking), reserving(walking), stealingpedestrian(walking, stealing, standing, fighting, talking, attacking, picking_up), badparking(walking), crash(walking, running, picking_up), stealingcar(stealing), stealinginside(running, stealing, walking), suspicious(talking, walking)

Drone-Action	13	240	clapping (10), jogging_f_b (20), jogging_side (20), running_f_b(20), running_side (20), walking_f_b (20) , walking_side (20) , hand_waving(10)	boxing (20), hitting_bottle (20), hitting_stick (20), kicking (20), stabbing (20)	boxing, hand_waving, clapping, hitting_bottle, hitting_stick, kicking, stabbing, jogging_f_b, jogging_side ,running_f_b,running_side,walking_f_b, walking_side
Okutama-Action	13	43	handshaking (24), hugging (23), reading(39), drinking(23), pushing /pulling(HOI) (21), carrying(30), calling(37), running(25), walking(43), lying(23), sitting(43), standing(43)	pushing/pulling (HHI) (22)	pushing/pulling (HHI), running, walking, drinking, sitting, handshaking, standing, reading, carrying, hugging, calling
NEC-Drone	16	2,079	walk (271), run (118), jump (120), pick up a backpack and go (113), leave a backpack and go (114), sit on a chair (116), talk on a mobile phone (112), drink water from a bottle (111), throw something (163), pick up a small object (141), shake hands (104), hug (107), exchange a backpack (100), walk toward each other and stay (145; walking, standing), stand together leave (146; walking, standing)	push a person (98)	push a person, walk, walk toward each other and stay (walking, standing),, run, jump, sit on a chair, talk on a mobile phone, drink water from a bottle, throw something, shake hands, stand together leave (walking, standing) , hug, pick up a small object, pick up a backpack and go, exchange a backpack
UAV Human	155	22,476	drink (175), eat snacks (174), brush hair (172), drop something (171), pick up something (169), throw away something (173), sit down (170), stand up (170), applaud (171), read (169), write (170), put on a coat (169), take off a coat (168), put on glasses (168), take off glasses (168), put on a hat (170), take off a hat (169), throw away a hat (169), cheer (169), wave hands (169), reach into pockets (166), jump on single leg	punch with fists (168), kick aside (165), kick backward (165), stagger (166), punching someone (113), kicking someone (111), pushing someone (110), steal something from other's pocket (110), rob something from someone (112), hit someone with something (113), threat some with a knife (111), bump into someone (113), hold someone hostage (113), threat someone with a gun (105), drag someone (101), stab someone with a knife (104), kick something (169), chest discomfort (165), pick a lock (106)	drink, drink a toast(drinking), punching someone, punch with fists, kicking someone, kick something , kick aside, kick backward , rob something from someone, pushing someone, sit down, stand up, applaud, make a phone call, pick up something , shake hands, read, hit someone with something, run, stab someone with a knife, walk toward someone, walk away from someone , walk side by side, walk, throw away something, throw away a hat, throw a frisbee, throw litter, steal something from other's pocket, jump on single leg, jump on two legs, carry a carrying pole, hug, exchange something with someone, all clear, have command

			(169), jump on two legs (169), make a phone call (171), play with cell phones (169), point somewhere (170), look at the watch (168), rub hands (167), bow (171), shake head (167), salute (170), cross palms together (168), cross arms in front to say no (170), wear headphones (166), take off headphones (167), make a shh sign (166), touch the hair (165), thumb up (168), thumb down (165), make an ok sign (167), make an ok sign (167), make a victory sign (168), figure snap (166), open the bottle (165), smell (165), squat (167), apply cream to face (166), apply cream to hands (167), grasp a bag (166), put down a bag (164), put something into a bag (165), take something out of a bag (164), open a box (165), move a box (166), put up hands (166), put hands on hips (168), wrap arms around (162), shake arms (164)		,hover, land, land at designated locations, move forward, move left, move right, ascend, descend, not clear, decelerate, call for help pick a lock, dig a hole, drag someone, threat some with a knife, hold someone hostage, threat someone with a gun

TABLE VII. HUMAN ACTIVITIES ANALYSIS OF THE UAV DATASETS (CONT.)

UAV Dataset	No. of actions	No. of videos	Normal Behavior	Abnormal Behavior	Common Activities
UAV Human	155	22,476	step on the spot walk (166), cough (168), sneeze (168), yawn (167), blow nose (166), headache (166), backache (164), neck-ache (163), vomit (164), use a fan (165), stretch body (165), point someone (113), hug (103), give something to someone (106), shake hands (107), walk toward someone (111), walk away from someone (111), walk side by side (109), high five (113), drink a toast (112; clap the bottle,drinking), move something with someone (110), take a phone for someone (115), stalk someone (111), whisper in someone's ear (107), exchange something with someone (109), lend an arm to support someone (106), rock-paper-scissors (109), hover (152), land (151), land at designated locations (151), move forward (150), move backward (149), move left (150) move right (149), ascend (149), descend (149), accelerate (149), decelerate (146), come over here (148), stay where you are (148), rear right turn (147), rear left turn (148), abandon landing		

			(147), all clear (147), not clear (147), have command (146), follow me (147), turn left (147), turn right (147), throw litter (143), dig a hole (152), mow (123), set on fire (144), smoke (150), cut the tree (152), fishing (147), pollute walls (157), wave a goodbye (102), comfort someone (102), sweep the floor (137), mop the floor (136), bounce the ball (138), shoot at the basket (138), swing the racket (137), leg pressing (138), escape (to survive) (133), call for help (140), wear a mask (137), take off a mask (135), bend arms around someone's shoulder (102), run (43), throw a frisbee (137), carry a carrying pole (142), walk (43), use a lever to lift something (135), close an umbrella (41), open an umbrella (42), slap someone on the back (113), chase someone (103)		
Aerial Gait	1	17	walking (17)	-	walking
UAV-GESTURE	13	119	all clear (11), have command (11), hover (7), land (7), landing direction (7), move ahead (11), move downward (7), move to left (11), move to right (11), move upward (7), not clear (11), slow down (11), wave off (7)	-	all clear, have command, ,hover, land, wave off, landing direction, move ahead, move to left, move to right, move upward, move downward, not clear, slow down
P-DESTRE	14	75	walking(73), running(0), standing(33), sitting(1), cycling(0), exercising(0), petting(0), talking over the phone(7), leaving bag(0), dating(0), trading(0)	offending(0), fall(0), fighting(0)	walking, standing, sitting, talking over the phone
UCF-ARG	17	473 (only UAV)	carrying(49), clapping(50), digging(50), jogging(49), open-close trunk(40), running(50), throwing(50), walking(50), waving(49), standing(2),picking_up(2),gesturing(2),tennis_swing(1),closing_trunk(1),opening_trunk(1), jump(1)	boxing(50)	boxing, carrying, clapping, digging, jogging, running, throwing, walking, waving, standing, gesturing, picking_up, jump
SAR-UAV	7	1880 images (5 videos)	after drop: standing (1293), walking(404), running(373) ,sitting(0), lying(0), handshake(2), and hand-waving(1014).	-	standing, walking, running, handshake, hand-waving.

TABLE VIII. UNIFIED PRECISE CATEGORIZATION OF COMMON HUMAN ACTIVITIES ON THE UAV DATASETS

Unified Precise Categorization		MDVD	Drone-Action	Okutama-Action	NEC-Drone	UAV Human	Aerial Gait	UAV-GESTURE	P-DESTRE	UCF-ARG	SAR-Drone
Abnormal	Fighting	Fighting(fighting) (HHI), Attack(attacking) (HHI), StealingPedestrian (fighting, attacking) (HHI)	Boxing(AO), Kicking(AO)	Pushing/Pulling (HHI)	Push A Person(HHI)	Pushing Someone(HHI), Punching Someone(HHI), Punch With Fists(AO), Kicking Someone(HHI), Kick Backward(AO), Kick Aside(AO),drag someone (HHI+HOI)	-	-	-	Boxing(AO)	-
	Robbery	Stealingpedestrian (stealing) (HHI+HOI), StealingInside(stealing) (HOI)	-	-	-	Steal Something From Other's Pocket (HHI+HOI), Rob Something From Someone(HHI+HOI)	-	-	-	-	-
	Vehicle Theft	StealingCar(stealing) (HOI), Attack(stealing) (HOI)	-	-	-	pick a lock(HOI)	-	-	-	-	-
	Assault	-	Hitting_Bottle(HOI), Hitting_Stick (HOI), Stabbing(HOI)	-	-	Hit Someone with Something(HHI+HOI), Stab Someone with A Knife(HHI+HOI), threat some with a knife (HHI+HOI), hold someone hostage(HHI+HOI), threat someone with a gun(HHI+HOI)	-	-	-	-	-
Normal	Walking	Normal(walking)(AO), BadParking(walking) (AO), Reserving(walking) (AO), Crash(walking) (AO), StealingInside(walking) (AO), StealingPedestrian (walking) (AO), Suspicious(walking) (AO)	walking_f_b(AO), walking_side (AO)	Walking (AO)	Walk (AO), Walk Toward Each Other And Stay(walking) (AO), Stand Together Leave(walking) (AO),	Walk (AO), Walk Toward Someone(AO), Walk Away From Someone(AO), Walk Side By Side(AO)	Walking (AO)	-	Walking (AO)	Walking (AO)	Walking (AO)
	Clapping	-	Clapping(AO)	-	-	Applaud(AO)	-	-	-	Clapping (AO)	-

	Digging	-	-	-	-	dig a hole(HOI)	-	-	-	Digging(HOI)	-
	Drinking	-	-	Drinking(HOI)	Drink Water From A Bottle(HOI)	Drink(HOI), Drink A Toast(drinking)(HOI)	-	-	-	-	-
	Sitting	-	-	Sitting(HOI)	Sit On A Chair(HOI)	Sit Down(HOI)	-	-	Sitting(HOI)	-	-
	Handshaking	-	-	Handshaking(HHI)	Shake Hands(HHI)	Shake Hands(HHI)	-	-	-	-	handshake(HHI)
	Standing	Normal(Standing)(AO), StealingPedestrian(standing) (AO)	-	Standing(AO)	Stand Together Leave (standing) (AO), Walk Toward Each Other And Stay(standing) (AO)	Stand Up(AO)	-	-	Standing (AO)	standing(AO)	Standing(AO)
	Reading	-	-	Reading(HOI)	-	Read(HOI)	-	-	-	-	-

TABLE IX. UNIFIED PRECISE CATEGORIZATION OF COMMON HUMAN ACTIVITIES ON THE UAV DATASETS (CONT.)

Unified Precise Categorization		MDVD	Drone-Action	Okutama-Action	NEC-Drone	UAV Human	Aerial Gait	UAV-GESTURE	P-DESTRE	UCF-ARG	SAR-Drone
Normal	Running	Attack(running)(AO),Crash(running)(AO), StealingInside(running)(AO)	Running (AO), Jogging (AO)	Running (AO)	Run(AO)	Run(AO)	-	-	-	Running (AO), Jogging(AO)	Running(AO)
	Jumping	-	-	-	Jump(AO)	Jump On Single Leg(AO), Jump on Two Legs(AO)	-	-	-	Jump(AO)	-
	Talking	Normal(Talking)(HHI),Fighting(talking)(HHI), StealingPedestrian(talking)(HHI), Suspicious(talking)(HHI)	-	Calling (HOI)	Talk On A Mobile Phone (HOI)	Make A Phone Call(HOI)	-	-	Talking Over the Phone(HOI)	-	-
	Throwing	-	-	-	Throw Something(HOI)	Throw Away Something(HOI), Throw Away A Hat(HOI), Throw A Frisbee(HOI), Throw Litter(HOI)	-	-	-	Throwing (AO)	-
	Carrying	-	-	Carrying (HOI)	-	Carry A Carrying Pole(HOI)	-	-	-	Carrying(HOI)	-
	Hugging	-	-	Hugging (HHI)	Hug(HHI)	Hug(HHI)	-	-	-	-	-
	Pick-Up	Crash(picking_up)(HOI), StealingPedestrian(picking_up)(HOI)	-	-	Pick Up a Small Object(HOI),	Pick Up Something(HOI)	-	-	-	picking_up(HOI)	-

					Pick Up a Backpack and Go(HOI)						
	Exchange Something	-	-	-	Exchange A Backpack(HHI+HHOI)	Exchange Something with Someone(HHI+HOI)	-	-	-	-	-
	Gesture Signal	All Clear	-	-	-	-	All Clear	-	All Clear	-	-	-
		Have Command	-	-	-	-	Have Command	-	Have Command	-	-	-
		Hover	-	-	-	-	Hover	-	Hover	-	-	-
		Land	-	-	-	-	Land	-	Land	-	-	-
		Landing Direction	-	-	-	-	Land at Designated Locations	-	Landing Direction	-	-	-
		Move Ahead	-	-	-	-	Move Forward	-	Move Ahead	-	-	-
		Move To Left	-	-	-	-	Move Left	-	Move To Left	-	-	-
		Move To Right	-	-	-	-	Move Right	-	Move To Right	-	-	-
		Move Upward	-	-	-	-	Ascend	-	Move Upward	-	-	-
		Move Downward	-	-	-	-	Descend	-	Move Downward	-	-	-
		Not Clear	-	-	-	-	Not Clear	-	Not Clear	-	-	-
		Slow Down	-	-	-	-	Decelerate	-	Slow Down	-	-	-
		Hand Waving	-	Hand Waving	-	-	call for help	-	Wave Off	-	Waving, gesturing	hand-waving

TABLE X. DISTINCT ABNORMAL ACTIONS ON THE UAV DATASETS

Distinct Abnormal Activities	MDVD	UAV Human
Falling	Falling (AO)	-
Crash	Crash (HHI+HOI)	
Stagger	-	Stagger (AO)
Chest Discomfort		Chest Discomfort (AO)
Bump into Someone		Bump into Someone (HHI)

Evaluation of the Impact of Cybersecurity Knowledge on the Prevention of Social Cybercrime Among University Students in Mexico, Colombia, and Peru

Yasmina Riega-Viru[1], Lainiver Mendoza Munar[2], Mario Ninaquispe-Soto[3], Kiara Nilupu-Moreno[4], Juan Luis Salas-Riega[5], Alfonso Renato Vargas-Murillo[6]*, Yolanda Pinto Bouroncle[7]
Universidad Privada del Norte, Lima, Peru[1, 3, 4, 6]
Universidad Cooperativa de Colombia, Cali, Colombia[2]
Pontificia Universidad Católica del Perú, Lima, Peru[5]
Universidad Peruana de Ciencias Aplicadas, Lima, Perú[7]

***Abstract*—Objectives: This study aims to evaluate the degree of cybersecurity knowledge and awareness among university students in Peru, Mexico, and Colombia, and to determine how these factors contribute to protection against social cybercrime. This cross-regional analysis represents a novel contribution by comparing cybersecurity preparedness across three Latin American countries, an underrepresented region in cybersecurity education research. Methods: A cross-sectional study was conducted using a 97-question survey that assessed both cybersecurity knowledge and practices. The study involved 809 university students from Peru, Mexico, and Colombia. Correlation analysis was performed to examine the relationship between cybersecurity knowledge and cybercrime prevention practices. Results: The analysis revealed a positive but low correlation (r=0.252) between cybersecurity knowledge and cybercrime prevention practices. Only 10.71% of preventive practices could be explained by acquired knowledge. Greater efficacy was observed in cyberstalking prevention compared to other forms of cybercrime. A significant gap was found between theoretical knowledge and practical application of cybersecurity, with only 44.6% of students receiving occasional information on the subject. Conclusions: This study highlights the urgent need to improve cybersecurity education in Latin American universities. The findings underscore the importance of integrating applied practices into cybersecurity curricula to strengthen students' ability to effectively counter cyber threats. Future educational initiatives should focus on bridging the gap between theoretical knowledge and practical application to enhance students' resilience against social cybercrime.**

Keywords*—Cybersecurity; social cybercrime; university students

I. Introduction

Technological advancement has transformed the way daily activities are conducted, especially since the Covid-19 pandemic, when dependence on internet access became crucial [1]. Although these advances have proven effective and beneficial in many aspects, it is essential to question whether the associated risks are being adequately considered. International organizations have expressed concern about the increased reliance on applications, online courses, banking services, and other digital resources. They emphasize the need for governments to implement effective strategies to combat cyber threats [2]. Most online vulnerabilities stem from insufficient cybersecurity knowledge among users [3]. For university students, the use of online technology is ubiquitous; however, they often lack awareness about the importance of cybersecurity and risky online behavior due to limited understanding of cyber security. Cybercrime is constantly evolving; for example, an emerging crime is "social cybercrime," which relates to online social interactions and represents the transposition of traditional conflicts to a digital environment [4]. First-world countries have implemented various mechanisms to prevent user vulnerability, such as raising awareness among high school and university students [5]; however, in Latin America, the same attention has not been paid to this problem. This research addresses a critical gap by conducting a comparative analysis across three Latin American countries—Peru, Mexico, and Colombia—to understand regional patterns in cybersecurity awareness and preparedness. Unlike previous single-country studies, this multi-national approach provides insights into shared vulnerabilities and educational needs across the region, contributing to the development of context-specific prevention strategies for social cybercrime in Latin American higher education institutions. Therefore, the question arises: What is the degree of knowledge and awareness about cybersecurity among university students in Peru, Mexico, and Colombia as protective measures against social cybercrime? The remainder of this paper is organized as follows: Section II reviews related work on cybersecurity awareness and social cybercrime; Section III describes the methodology employed in this study; Section IV presents the results of our analysis; Section V discusses the findings and their implications; and Section VI concludes with recommendations for future research and practice.

II. Related Work

A systematic literature review was conducted examining studies published between 2018 and 2023. This review revealed that it is necessary to reinforce the knowledge, attitude, and behavior of university students in cybersecurity, as few countries implement preventive measures in this group [6]. Another descriptive, correlational, and cross-sectional study with 48 female high school students, focused on social

*Corresponding author.

cybercrime, showed that cybersecurity is effective in preventing cyberharassment and cyberbullying [7]. Also, Abdukadir Ahmed and colleagues [8] evaluated cybersecurity awareness in 250 undergraduate and graduate students from five universities. They found that virus and phishing attacks are the most common problems, therefore, they recommended continuous education and supervision in cybersecurity.

The constant evolution of cyber threats, such as ransomware attacks, highlights the growing importance of cybersecurity [9]. This has become essential to prevent data breaches, inactivity, and intellectual property theft [10]. Recent research highlights the growing dependence of individuals, organizations, and countries on digital tools and networks. However, this trend has intensified cybercriminal activities, which in turn highlights the need to continuously improve cybersecurity measures in the face of emerging threats [11]. In a related study, Almansoori and others [12] discovered that the Protection Motivation Theory (PMT) is predominant in most studies on cybersecurity behavior, noting that, of 39 articles analyzed, 56% focused on the organizational level, while research at the individual level is still in an initial stage.

Cybersecurity, crucial for protecting systems, networks, and data against digital attacks and ensuring the confidentiality, integrity, and availability of information, requires a comprehensive understanding of its environment, resources, and guidelines [13, 14, 15]. Currently, organizations must establish adequate levels of access and security gateways to handle information securely. The complexity of cybersecurity suggests a holistic approach, evidenced in the analysis of the legal implications of technologies such as blockchain [16].

Moreover, the theory of planned behavior [17] highlights the influence of perceived control and subjective norms on personal actions, relevant to information management and risks in cyberspace. The protection motivation theory [18, 19] complements this approach by predicting the intention to engage in protective practices, integrating risk perception and evaluation of the effectiveness of protective measures.

In the university context, it is vital to develop educational strategies that integrate technical knowledge, understanding of student motivations and behaviors in the face of digital threats, and promote effective security practices. This holistic approach contributes to training professionals capable of navigating cyberspace safely, thus strengthening the digital resilience of society.

There are various models designed to measure and improve cybersecurity knowledge. One of them is the "citizen-centered cybersecurity model" developed by Mahlangu et al. [20]. This model focuses on assessing citizen awareness, threat evaluation, coping mechanisms, attitudes, and behavioral intentions. The purpose is to promote safe behaviors in the digital realm, emphasizing that adequate knowledge is fundamental to understanding and effectively responding to cyber threats. This model integrates concepts from the Protection Motivation Theory (PMT).

On the other hand, the Qualification Level Awareness Measurement Model, proposed by Erol and Sagiroglu [21], offers a framework to direct attention to crucial aspects of a situation, integrate perceived information, and project future system states. This model emphasizes the importance of situational awareness, mental models, and information processing, essential aspects for understanding and improving cybersecurity practices. Applying these principles allows individuals to evaluate cyber threats more effectively, understand system vulnerabilities, and make informed decisions to mitigate risks.

Furthermore, Tirumala et al. [22] propose a tiered structure through the Qualification Level Awareness Measurement Model, arguing that situational awareness or the situation model reflects the current state of the mental model. This model captures not only system parameters but also their interrelationships, facilitating a meaningful understanding of the system's state and its impact on future events. By employing principles of the goal-driven processing model, cybersecurity professionals can make more effective decisions when responding to security incidents or implementing preventive measures, improving cybersecurity posture and risk management.

For the development of social cybercrime, the theory proposed by Miró Linares [4] is taken as a foundation. According to him, the complexity of characterizing the authors of social cybercrimes is due to the diversity of motivations and actions they encompass. This category includes crimes with different purposes, from sexual motivations to verbal aggression on social networks and forums. He points out that minors can not only be victims but also perpetrators of these crimes, actively participating in activities such as sexting, insults, slander, and cyberbullying. It is mentioned that the profiles of social cybercrime perpetrators are as varied as the types of crimes themselves, and that their characteristics may be similar to those who commit similar crimes in physical space. However, cyberspace modifies the risk environment and, therefore, can alter the profile of offenders. As can be noted, there is a need for a deeper analysis of how cyberspace changes the profile of perpetrators of crimes such as cybergrooming, cyberstalking, and cyberbullying.

Cyberbullying is a phenomenon that merges technology and school bullying, using platforms such as social networks and instant messaging to repeatedly harass individuals, causing significant psychological damage to young people and university students [23, 24, 25]. Information and Communication Technologies (ICT) facilitate both communication and amplified forms of bullying [26, 27], also affecting university students, who require adjusted prevention strategies [28]. Cyberbullying is recognized by the European Union as one of the main online risks, driving measures to protect young people [29], while research such as that of Faucher et al. [30] suggests that impacts may vary by gender.

Cyberstalking, an evolution of traditional stalking, involves the use of technology to pursue and threaten, characterized by persistence and methods such as threatening messages and identity theft [31, 32]. This phenomenon is debated among experts on whether it constitutes an extension of conventional harassment or a distinct problem [33]. Research indicates that characteristics such as low self-control and bad social influences predispose to cyberstalking [34], and studies in Italy

reveal that victims can experience severe psychological consequences, especially those exposed to multiple forms of harassment [35].

Sexting, the exchange of erotic images through electronic devices, is common among university students, especially under pandemic circumstances [36, 37]. Although generally consensual, it can lead to risky practices such as non-consensual content publication, highlighting the importance of cybersecurity education to prevent abuse and blackmail [38, 39].

Online grooming is another severe form of online abuse where adults seek to manipulate minors to sexually abuse them, differing from sexting in its intentions and consequences [40, 41]. A study revealed that a significant proportion of Spanish university students has been affected by this phenomenon, with a notable impact on women and non-binary individuals [42]. Research also points to the influence of pornography consumption on vulnerability to grooming [43].

This analysis underscores the urgent need to implement robust educational and protection measures in academic settings to combat cybercrime and ensure a safe environment for students. Therefore, the objective of this study is to evaluate the degree of knowledge and awareness about cybersecurity among university students in Peru, Mexico, and Colombia, and determine how these factors contribute to protection against social cybercrime.

III. Methodology

A basic research of descriptive nature with a quantitative approach was conducted; of non-experimental cross-sectional design and correlational explanatory level, with the purpose of establishing the relationship between variable "X" representing the Level of Knowledge in Cybersecurity and variable "Y", representing the Degree of Prevention in Cybercrime.

For data collection, a questionnaire composed of 66 questions related to variable "X" and 31 questions for variable "Y" was used. The dimensions evaluated for each variable included the Level of Knowledge in aspects such as Online grooming, Cyberstalking, Cyberbullying, and Sexting for variable "X", and for variable "Y", the Degree of Prevention and exposure. To validate and ensure the reliability of the instrument, a pilot test was conducted, and the Cronbach's Alpha coefficient was calculated.

TABLE I. Reliability Statistics of the Instrument

Cronbach's Alpha	N of elements
0.962	98

Table I shows the reliability of the applied instrument, amounting to 0.962, which indicates that the applied instrument is highly reliable.

The analysis theory for the cybersecurity variable was based on KAB model [6], for the social cybercrime variable it was based on the typology of cybercrime proposed by Miró Linares.

The population consisted of university students from three Latin American universities (Mexico, Colombia, and Peru). The study was conducted with a sample of 809 students, who voluntarily responded to the questionnaire through a Google form.

The data exported to Excel were processed in the SPSS statistical package version 28. To identify the characteristics of students according to origin, age, study cycle; bivariate statistical analysis of association was applied, data normality analysis was performed, and considering the nature of the variables, Spearman's Rho statistical test was used to measure the correlation between them and their dimensions.

IV. Results

The processed data correspond mostly (57%) to students in cycles II, IV, VI, and VII (Fig. 1)

Fig. 1. Study cycle of students.

Fig. 2 shows that the largest age range of participating students is between 15 and 25 years old with 71%, and only 29% corresponds to students between 26 to 64 years old. Dispersion measures show that the mean age is 25 years with a standard deviation of 8.733 years.

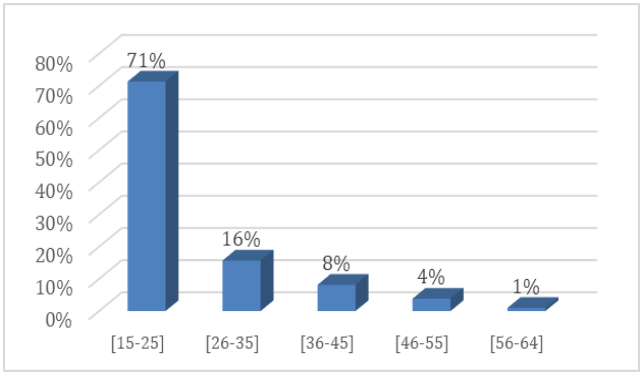


Fig. 2. Age range of participating students.

As shown in Fig. 3, 58% of the surveyed students come from the Universidad Privada del Norte Lima campus, 15% from UAQ Mexico, 13% from the Universidad Tecnológica del Perú, and the remaining 14% from UCC Colombia, UNIFÉ Lima, UPC Lima. This distribution reflects the multi-institutional and cross-national character of the study sample.

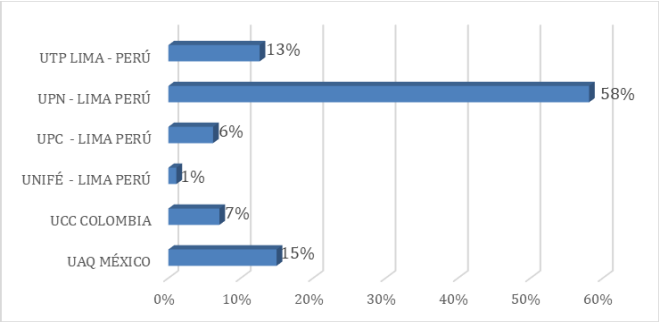


Fig. 3. Universities of origin.

Table II shows that 44.6% of students occasionally receive or have had knowledge about Cybersecurity, with UPN-Lima having the highest level on the subject (27.6%); likewise, 40.3% of students indicate that they have rarely evidenced or put into practice prevention in cybercrime offenses, with UPN-Lima again having the highest percentage with 24.6%.

The data normality analysis for the variables and their dimensions indicates a significance p-value = 0.000 < 0.001, so it is concluded that the behavior of the data does not follow a normal distribution, therefore, it is necessary to apply non-parametric tests for the validation of assumptions (Table III).

TABLE II. CROSS-TABULATION ON LEVEL OF KNOWLEDGE IN CYBERSECURITY AND DEGREE OF PREVENTION IN CYBERCRIME BY UNIVERSITY

		University						Total
		UAQ México	UCC Colombia	UNIFÉ Lima Peru	UPC Lima Peru	UPN Lima Peru	UTP Lima Peru	
X: Level of Knowledge in Cybersecurity	Never	0.5%	0.5%		0.2%	1.7%	1.5%	4.4%
	Rarely	6.8%	2.2%	0.2%	2.8%	20.1%	6.7%	38.9%
	Ocassionally	6.9%	3.3%	0.7%	2.5%	27.6%	3.6%	44.6%
	Frequent	0.7%	0.7%	0.1%	0.6%	7.4%	0.6%	10.3%
	Very Common		0.2%			1.2%	0.2%	1.7%
Y: Degree of prevention in cybercrime	Never	6.7%	2.2%	0.2%	1.6%	19.5%	7.0%	37.3%
	Rarely	5.3%	3.1%	0.2%	3.2%	24.6%	3.8%	40.3%
	Ocassionally	2.3%	1.1%	0.5%	1.0%	10.4%	1.4%	16.7%
	Frequent	0.4%	0.5%	0.1%	0.2%	2.7%	0.4%	4.3%
	Very Common	0.2%	0.1%		0.1%	0.9%		1.4%
Total		15.0%	7.0%	1.1%	6.2%	58.1%	12.6%	100.0%

TABLE III. DATA NORMALITY ANALYSIS

	Kolmogorov-Smirnov[a]		
	Statistical	gl	Sig.
D1X: Learning	0.213	809	0.000
D2X: Knowledge	0.223	809	0.000
D3X: Practice	0.242	809	0.000
D4XLearning_ Online Grooming	0.224	809	0.000
D5X: Knowledge_Arrangement Online	0.218	809	0.000
D6X: Practice_Online Arrangement	0.265	809	0.000
D7X: Learning_Cyberstalking	0.204	809	0.000
D8X: Knowledge_Cyberbullying	0.205	809	0.000
D9X: Practice_Cyberstalking	0.228	809	0.000
D10X: Learning_Cyberbullying	0.213	809	0.000
D11X: Knowledge_Cyberbullying	0.223	809	0.000
D12X: Practice_Cyberbullying	0.249	809	0.000
D13X: Learning_Sexting	0.201	809	0.000
D14X: Knowledge_Sexting	0.192	809	0.000
D15X: Practice__Sexting	0.245	809	0.000
D1Y_Victimization	0.224	809	0.000
D2Y_ GROOMING Exhibition	0.360	809	0.000
D3Y_ GROOMING Victimization	0.208	809	0.000
D4Y_CYBERSTALKING Exhibition	0.237	809	0.000
D5Y_Victimization CYBERBULLYING	0.247	809	0.000
D6Y_CYBERBULLYING Victimization	0.203	809	0.000
D7Y_SEXTING Exhibition	0.383	809	0.000
X: Level of cybersecurity knowledge	0.233	809	0.000
Y: Degree of prevention of cybercrime	0.242	809	0.000

TABLE IV. ANALYSIS OF CORRELATIONS BETWEEN THE STUDY VARIABLES

		Y: Degree of prevention of cybercrime
X: Level of cybersecurity knowledge	Spearman's Rho correlation coefficient	,252**
	Sig. (bilateral)	0.000 (p-valor<0.01)

Table IV shows the existing correlation between variable X: Level of knowledge in cybersecurity and variable Y: Degree of prevention in cybercrime, evidencing the existence of a correlation (p-value = 0.000 < 0.01), positive low equal to the Spearman's Rho coefficient 0.252. This statement is corroborated by Fig. 4 where the correlation points between both variables are shown, where the coefficient of determination $R^2 = 0.1071$ indicates that only 10.71% of the degree of prevention in criminality is given by the level of knowledge in cybersecurity.

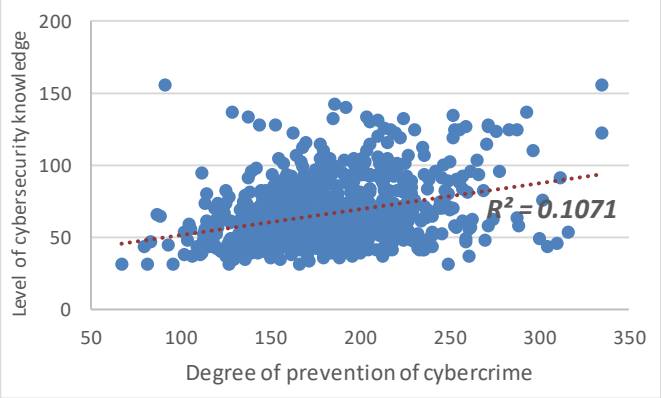


Fig. 4. Scatter plot of correlation between the level of knowledge in cybersecurity and the degree of prevention in cybercrime.

Table V shows the existing correlations between the dimensions of variables X and Y:

- There is a low positive correlation (r=0.134) between Cybersecurity Learning and Victimization in Cybercrime.
- There is a low positive correlation (r=0.239) between Cybersecurity Practice and Victimization in Cybercrime.

Regarding Grooming, it is evident that its Learning has a low positive correlation (r=0.095) with its Victimization; Knowledge and Practice have a low positive correlation with its Exposure (r=0.133, r=0.159) respectively, likewise they have a low positive correlation with its Victimization (r=0.114, r=0.223) respectively.

TABLE V. ANALYSIS OF CORRELATIONS BETWEEN THE STUDY VARIABLES

			Y						
			D1Y Victimization	**D2Y Groomin Exhibition**	**D3Y Victimization Grooming**	**D4Y Cyberstalking Exhibition**	**D5Y Cyberstalking Victimization**	**D6Y Cyberbulluing Victimization**	**D7Y Sexting Exhibition**
X	D1X: Learning	Correlation coefficient	,134**	,069*	,111**	,194**	,143**	,132**	0.059
		Sig. (bilateral)	0.000	0.050	0.002	0.000	0.000	0.000	0.096
	D2X: Knowledge	Correlation coefficient	0.065	,071*	0.044	,173**	,080*	0.053	0.017
		Sig. (bilateral)	0.063	0.042	0.215	0.000	0.024	0.135	0.627
	D3X: Practice	Correlation coefficient	,239**	,155**	,165**	,233**	,210**	,156**	,158**
		Sig. (bilateral)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	D4X: Online Grooming Learning	Correlation coefficient	,184**	0.024	,095**	,193**	,131**	,123**	0.060
		Sig. (bilateral)	0.000	0.489	0.007	0.000	0.000	0.000	0.088
	D5X: Online Grooming Knowledge	Correlation coefficient	,157**	,133**	,114**	,190**	,157**	,120**	,084*
		Sig. (bilateral)	0.000	0.000	0.001	0.000	0.000	0.001	0.017
	D6X: Online Grooming Practice	Correlation coefficient	,344**	,159**	,223**	,277**	,280**	,207**	,178**
		Sig. (bilateral)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	D7X: Cyberstalking Learning	Correlation coefficient	,118**	,074*	,082*	,220**	,103**	,108**	0.040
		Sig. (bilateral)	0.001	0.036	0.019	0.000	0.003	0.002	0.251
	D8X: Cyberstalking Knowledge	Correlation coefficient	,111**	0.038	,076*	,173**	,097**	,095**	0.005
		Sig. (bilateral)	0.002	0.278	0.030	0.000	0.006	0.007	0.896
	D9X: Cyberstalking Practice	Correlation coefficient	,424**	,166**	,263**	,287**	,345**	,252**	,202**
		Sig. (bilateral)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	D10X: Learning Cyberbullying	Correlation coefficient	,134**	,069*	,111**	,194**	,143**	,132**	0.059
		Sig. (bilateral)	0.000	0.050	0.002	0.000	0.000	0.000	0.096
	D11X: Cyberbullying Knowledge	Correlation coefficient	0.065	,071*	0.044	,173**	,080*	0.053	0.017
		Sig. (bilateral)	0.063	0.042	0.215	0.000	0.024	0.135	0.627
	D12X: Cyberbullying Practice	Correlation coefficient	,344**	,167**	,220**	,235**	,283**	,213**	,175**
		Sig. (bilateral)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	D13X: Learning Sexting	Correlation coefficient	,305**	,171**	,216**	,312**	,305**	,232**	,161**
		Sig. (bilateral)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	D14X: Sexting Knowledge	Correlation coefficient	,154**	,090*	,088*	,212**	,148**	,113**	0.050
		Sig. (bilateral)	0.000	0.011	0.013	0.000	0.000	0.001	0.156
	D15X: Sexting Practice	Correlation coefficient	,422**	,198**	,301**	,301**	,337**	,262**	,221**
		Sig. (bilateral)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	X	Correlation coefficient	,285**	,148**	,174**	,267**	,249**	,189**	,141**
		Sig. (bilateral)	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Regarding Cyberstalking, it is evident that its Learning has a low positive correlation with its Exposure and Victimization (r=0.220, r=0.103) respectively; Knowledge about this crime has a low positive correlation with Exposure and Victimization (r=1.73, r=0.097) respectively, finally its Practice has a low positive correlation with Exposure and Victimization (r=0.287, r=0.345) respectively. Regarding Cyberbullying, it is evident that Practice in its knowledge has a low positive correlation (r=0.213) with the Victimization of this crime.

Regarding Sexting, it is evident that Learning and Practice in the recognition of this crime have a low positive correlation (r=0.161, r=0.221) with its Exposure, respectively.

Finally, variable X (Level of knowledge in cybersecurity) has a low positive correlation with Exposure and Victimization of the crimes: Grooming, Cyberstalking, Cyberbullying, and Sexting, which leads to the need to improve the mechanisms and procedures of knowledge about cybersecurity, which improve this correlation and thus avoid exposure and victimization of these crimes.

V. Discussions

This study on the knowledge and awareness of cybersecurity of social cybercrime among university students from Mexico, Colombia, and Peru reveals significant findings that indicate the need to improve cybersecurity education for this demographic group.

It is observed that the studied university population presents a diversity in terms of experience and exposure to knowledge about cybersecurity. Although the average age is 25 years, with a standard deviation of 8.733 years, it is important to note that 71% of the participants are in the range of 15 to 25 years, while only 29% corresponds to students between 26 to 64 years. This distribution suggests that a significant proportion of students are younger than average.

Although it could be expected that this population has had time to be exposed to knowledge about cybersecurity, the data indicates that only 44.6% of students occasionally receive or have had knowledge on the subject. Furthermore, the low elevation (r=0.252) between the level of knowledge in cybersecurity and the degree of prevention of cybercrime suggests that age and exposure time do not necessarily translate into greater awareness or effective practice in relation to social cybercrime in its various manifestations.

A. Gap Between Theoretical Knowledge and Practical Application of Cybersecurity

The data obtained in this study reveal a concerning gap between theoretical knowledge of cybersecurity and its practical application by university students. Despite having a relatively high level of theoretical awareness about cyber threats, many students do not effectively apply this knowledge to protect themselves. This is reflected in the modest correlations between the level of knowledge and the degree of prevention of cybercrime, indicating that knowledge alone is not sufficient to guarantee safe online behaviors [44, 45]. The KAB model [6] suggest that cybersecurity education must go beyond the simple transfer of information and focus on the formation of practical skills and motivation to effectively apply this knowledge.

B. Variability in Awareness and Prevention of Different Types of Cybercrime

The study also shows significant variability in awareness and prevention among different types of cybercrimes. For example, awareness about cyberbullying seems to be higher than about online grooming, which may be influenced by the frequency and visibility of prevention campaigns focused on school bullying. However, the lack of awareness and effective prevention in less discussed crimes, such as cyberstalking and sexting, suggests the need for more inclusive and diversified educational approaches that address all forms of cybercrime [4, 42], ensuring that students are equipped to recognize and respond to a wider range of threats.

C. International Collaboration and Multidisciplinary Approach

International collaboration and a multidisciplinary approach are crucial aspects for effectively combating cybercrime. This study, which encompasses universities from Peru, Mexico, and Colombia, underscores the importance of this perspective. The variability in legislation and security practices between different countries demands greater cooperation and coordination between governmental entities, educational institutions, and private sectors at a global level. Furthermore, integrating approaches from multiple disciplines, such as psychology, law, and information technology, can offer more comprehensive perspectives and more effective solutions to cybersecurity challenges, especially in academic and youth environments [17, 18].

VI. Conclusion

This study provides empirical evidence of the cybersecurity knowledge and preventive practices among 809 university students across Peru, Mexico, and Colombia. The findings reveal a positive but weak correlation ($r=0.252$, $p<0.01$) between cybersecurity knowledge and cybercrime prevention practices, with only 10.71% of variance in preventive behaviors explained by knowledge levels ($R^2=0.1071$). These quantitative results indicate that 44.6% of students receive only occasional cybersecurity training, and 40.3% rarely implement preventive measures against social cybercrime. Among the four types of social cybercrime examined, cyberstalking prevention demonstrated the highest correlation with knowledge (r=0.287 for practice-exposure and r=0.345 for practice-victimization), while grooming showed the weakest associations.

A. Theoretical and Practical Implications

This research contributes to cybersecurity education theory by empirically validating the gap between theoretical knowledge and behavioral application in the Latin American context, extending the Protection Motivation Theory (PMT) and the Knowledge-Attitude-Behavior (KAB) model to the domain of social cybercrime prevention. The multi-national comparative approach reveals that cybersecurity challenges are regionally consistent across Mexico, Colombia, and Peru, suggesting that coordinated educational interventions could be effective across Latin America.

From a practical standpoint, these findings have direct implications for higher education institutions. Universities should redesign cybersecurity curricula to emphasize hands-on training, simulation-based learning, and real-world case studies rather than solely theoretical instruction. Educational programs must specifically address the four dimensions of social cybercrime—grooming, cyberstalking, cyberbullying, and sexting—with particular attention to grooming prevention, which showed the weakest knowledge-practice correlation. Furthermore, institutions should implement mandatory, continuous cybersecurity awareness programs integrated throughout the academic curriculum rather than optional or occasional workshops. Policymakers in Latin American countries should consider establishing regional cybersecurity education standards and sharing best practices to strengthen collective digital resilience.

B. Research Limitations

This study presents several limitations that should be acknowledged. First, the cross-sectional design captures only a snapshot of cybersecurity knowledge and practices, preventing causal inferences or the examination of temporal changes in behavior. Second, the convenience sampling method, while yielding a substantial sample size (n=809), may not fully represent the diversity of university students across all Latin American countries, as the study focused on three specific nations. Third, self-reported measures are susceptible to social desirability bias, potentially leading participants to overestimate their cybersecurity knowledge or underreport risky online behaviors. Fourth, the study did not account for potential confounding variables such as socioeconomic status, prior technology education, or individual differences in digital literacy that may moderate the knowledge-practice relationship. Finally, the instrument, while demonstrating high reliability (α=0.962), was not validated across different cultural contexts, which may affect its applicability to other regions or populations.

C. Future Research Directions

Based on these findings and limitations, we propose three specific directions for future research. First, longitudinal studies should be conducted to track changes in cybersecurity knowledge and preventive behaviors over time, particularly following targeted educational interventions. Such studies could employ pre-test/post-test experimental designs to establish causal relationships between specific teaching methodologies and behavioral outcomes. Second, future research should expand geographically to include other Latin American countries and develop cross-cultural comparisons with other global regions to identify universal versus culture-specific factors influencing cybersecurity behaviors. This expansion would enable the development of globally informed yet locally adapted prevention strategies. Third, researchers should investigate the psychological and social factors that mediate or moderate the knowledge-practice gap, such as risk perception, self-efficacy, peer influence, and institutional support. Mixed-methods approaches combining quantitative surveys with qualitative interviews could provide deeper insights into the barriers preventing students from translating knowledge into protective action. Additionally, future studies should explore the effectiveness of innovative pedagogical approaches, such as gamification, virtual reality simulations, and peer-to-peer learning, in enhancing practical cybersecurity skills among university students.

REFERENCES

[1] C. Cardona-Londoño, M. Ramirez-Sanchez, and E. Rivas-Trujillo, "Educación Superior en un mundo virtual, forzado por la pandemia del Covid 19," Revista Espacios, vol. 41, no. 35, pp. 1–14, 2020.

[2] Organizaction of American States, "Ciberseguridad Marco NIST," 2019. [Online]. Available: https://www.oas.org/es/sms/cicte/docs/OEA-AWS-Marco-NIST-de-Ciberseguridad-ESP.pdf

[3] Organization of American States, "Cybersecurity Education Future Planning through Workforce Development," 2020. [Online]. Available: https://www.oas.org/es/sms/cicte/docs/White-Paper-Cybersecurity-Education.pdf

[4] F. Miró Linares, Cibercrimen Fenomenología y criminología de la delincuencia en el ciberespacio. Marcial Pons Ediciones Jurídicas y Sociales, 2012.

[5] M. A. Alqahtani, "Factors Affecting Cybersecurity Awareness among University Students," Applied Sciences, vol. 12, no. 5, p. 2589, 2022, doi: 10.3390/app12052589.

[6] K. Nilupú-Moreno, J. L. Salas-Riega, M. Ninaquispe-Soto, and Y. Riega-Virú, "Cybersecurity in University Students: A Systematic Review of the Literature," in Lecture Notes in Networks and Systems, 2024, pp. 315–332. doi: 10.1007/978-981-99-7886-1_27.

[7] Y. Riega-Virú, K. Nilupu-Moreno, J. L. Salas-Riega, and M. Ninaquispe-Soto, "Knowledge of cybersecurity against social cybercrime of female high school students," in Proceedings of the 2023 IEEE 3rd International Conference on Advanced Learning Technologies on Education and Research, ICALTER 2023, 2023. doi: 10.1109/ICALTER61411.2023.10372927.

[8] A. Abdukadir Ahmed, A. Hussein Elmi, A. Abdullahi, and A. Yahye Ahmed, "Cybersecurity awareness among university students in Mogadishu: a comparative study," Indonesian Journal of Electrical Engineering and Computer Science, vol. 32, no. 3, p. 1580, 2023, doi: 10.11591/ijeecs.v32.i3.pp1580-1588.

[9] H. Almonajid, "Fundamentals of Cyber Security," International Journal of Emerging Multidisciplinaries: Security, vol. 1, no. 1, 2023, doi: 10.54938/ijemds.2023.01.1.232.

[10] M. Mijwil, "The Significance of Machine Learning and Deep Learning Techniques in Cybersecurity: A Comprehensive Review," Iraqi Journal for Computer Science and Mathematics, pp. 87–101, 2023, doi: 10.52866/ijcsm.2023.01.01.008.

[11] A. Loishyn, S. Hohoniants, M. Tkach, M. Tyshchenko, N. Tarasenko, and V. Kyvliuk, "Development of the Concept of Cybersecurity of the Organization," TEM Journal, pp. 1447–1453, 2021, doi: 10.18421/TEM103-57.

[12] A. Almansoori, M. Al-Emran, and K. Shaalan, "Exploring the Frontiers of Cybersecurity Behavior: A Systematic Review of Studies and Theories," Applied Sciences, vol. 13, no. 9, p. 5700, 2023, doi: 10.3390/app13095700.

[13] B. Yadav, "Overview of Cyber Security," International Journal of Advanced Research in Science, Communication and Technology, pp. 489–492, 2022, doi: 10.48175/IJARSCT-3959.

[14] S. Chopra, H. Marwaha, and A. Sharma, "Cyber-Attacks Identification and Measures for Prevention," in Proceedings of the International Conference on Cybersecurity and Cybercrime (IC3), 2022, pp. 83–90. doi: 10.19107/CYBERCON.2022.11.

[15] M. Haripriya, "Cyber Security Unveiled: Trends and Protections in the Digital World," International Journal Of Scientific Research In Engineering And Management, vol. 07, no. 07, 2023, doi: 10.55041/IJSREM24720.

[16] L. Mendoza Munar, and A. Riascos Diaz, "Aspectos jurídicos de la tecnología Blockchain". Law, State & Telecommunications Review/Revista de Direito, Estado e Telecomunicações, vol. 13, no. 1

[17] I. Ajzen, "The theory of planned behavior," Organizational Behavior and Human Decision Processes, vol. 50, no. 2, pp. 179–211, 1991, doi: 10.1016/0749-5978(91)90020-T.

[18] R. W. Rogers, "Cognitive and physiological processes in fear appeals and attitude change: A revised theory of protection motivation," in Social psychophysiology: A sourcebook, 1983, pp. 153–176.

[19] I. Woon, G.-W. Tan, and R. Low, "A Protection Motivation Theory Approach to Home Wireless Security," in ICIS 2005 Proceedings, 2005.

[20] G. Mahlangu, C. Chipfumbu Kangara, and F. Masunda, "Citizen-centric cybersecurity model for promoting good cybersecurity behaviour," Journal of Cyber Security Technology, vol. 7, no. 3, pp. 154–180, 2023, doi: 10.1080/23742917.2023.2217535.

[21] S. E. Erol and S. Sagiroglu, "Awareness Qualification Level Measurement Model," in 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), 2018, pp. 107–112. doi: 10.1109/IBIGDELFT.2018.8625305.

[22] S. S. Tirumala, M. R. Valluri, and G. Babu, "A survey on cybersecurity awareness concerns, practices and conceptual measures," in 2019 International Conference on Computer Communication and Informatics (ICCCI), 2019, pp. 1–6. doi: 10.1109/ICCCI.2019.8821951.

[23] W. Lim, B. T. Lau, and F. M. A. Islam, "Cyberbullying Awareness Intervention in Digital and Non-digital Environment for Youth: Current Knowledge," Education and Information Technologies, vol. 28, no. 6, pp. 6869–6925, 2023, doi: 10.1007/s10639-022-11472-z.

[24] A. C. Beluce, K. L. de Oliveira, A. S. Ferraz, and L. da S. Almeida, "Cyberbullying and Motivation to Learn with Digital Technologies: Identification and Correlation," Psicologia: Teoria e Pesquisa, vol. 39, no. spe, 2023, doi: 10.1590/0102.3772e39nspe07.en.

[25] A. Bussu, M. Pulina, S.-A. Ashton, and M. Mangiarulo, "Exploring the impact of cyberbullying and cyberstalking on victims' behavioural changes in higher education during COVID-19: A case study," International Journal of Law, Crime and Justice, vol. 75, p. 100628, 2023, doi: 10.1016/j.ijlcj.2023.100628.

[26] L. Fernández, "Formación TIC (redes sociales, internet, ciberseguridad, big data, etc.) en casa, en el colegio, en la universidad y en la empresa: características, razón de ser y contenido," Revista Tecnología, Ciencia y Educación, vol. 12, no. 12, pp. 89–110, 2019.

[27] J. Kim, H. Song, and W. G. Jennings, "A Distinct Form of Deviance or a Variation of Bullying? Examining the Developmental Pathways and Motives of Cyberbullying Compared With Traditional Bullying in South Korea," Crime & Delinquency, vol. 63, no. 12, pp. 1600–1625, 2017, doi: 10.1177/0011128716675358.

[28] R. Ortega-Ruiz, R. Del Rey, and J. A. Casas, "Evaluar el bullying y el cyberbullying validación española del EBIP-Q y del ECIP-Q," Psicología Educativa, vol. 22, no. 1, pp. 71–79, 2016, doi: 10.1016/j.pse.2016.01.004.

[29] Euronews, "Lucha contra el ciberacoso: la UE estudia medidas para castigarlo por ley," Euronews, Jun. 17, 2023. [Online]. Available: https://es.euronews.com/2023/06/17/lucha-contra-el-ciberacoso-la-ue-estudia-medidas-para-castigarlo-por-ley

[30] C. Faucher, M. Jackson, and W. Cassidy, "Cyberbullying among University Students: Gendered Experiences, Impacts, and Perspectives," Education Research International, vol. 2014, pp. 1–10, 2014, doi: 10.1155/2014/698545.

[31] R. T. Gopalan, "Offending, victimization, forensic investigation, and prevention of cyberstalking," in Developing Safer Online Environments for Children: Tools and Policies for Combatting Cyber Aggression, 2019. doi: 10.4018/978-1-7998-1684-3.ch001.

[32] T. Begotti, M. A. Ghigo, and D. Acquadro Maran, "Victims of Known and Unknown Cyberstalkers: A Questionnaire Survey in an Italian Sample," International Journal of Environmental Research and Public Health, vol. 19, no. 8, p. 4883, 2022, doi: 10.3390/ijerph19084883.

[33] A. K. Gautam and A. Bansal, "Email-Based Cyberstalking Detection On Textual Data Using Multi-Model Soft Voting Technique Of Machine Learning Approach," Journal of Computer Information Systems, vol. 63, no. 6, pp. 1362–1381, 2023, doi: 10.1080/08874417.2022.2155267.

[34] C. D. Marcum, G. E. Higgins, and J. Nicholson, "I'm Watching You: Cyberstalking Behaviors of University Students in Romantic Relationships," American Journal of Criminal Justice, vol. 42, no. 2, pp. 373–388, 2017, doi: 10.1007/s12103-016-9358-2.

[35] T. Begotti and D. Acquadro Maran, "Characteristics of Cyberstalking Behavior, Consequences, and Coping Strategies: A Cross-Sectional Study in a Sample of Italian University Students," Future Internet, vol. 11, no. 5, p. 120, 2019, doi: 10.3390/fi11050120.

[36] L. A. Reed, M. P. Boyer, H. Meskunas, R. M. Tolman, and L. M. Ward, "How do adolescents experience sexting in dating relationships? Motivations to sext and responses to sexting requests from dating partners," Children and Youth Services Review, vol. 109, p. 104696, 2020, doi: 10.1016/j.childyouth.2019.104696.

[37] J. Mikova, "University Students' Risky Behavior In Online Environment During The Covid-19 Pandemic," in 6th EPIDAPO, International Conference on Applied Psychology and Educational Sciences, 2021, pp. 63–73. doi: 10.15405/epiceepsy.21101.5.

[38] A. Dodaj et al., "Through the Eyes of Young People: A Qualitative Study of Sexting Among Croatian and Bosnian and Herzegovinian College Students," Sexuality & Culture, vol. 26, no. 5, pp. 1885–1918, 2022, doi: 10.1007/s12119-022-09976-4.

[39] A. O. Dunmade, A. Tella, and U. D. Onuoha, "Awareness of Cyberethical Behavior Among Female Postgraduate Students in North Central Nigeria," Journal of Information Ethics, vol. 32, no. 2, pp. 122–136, 2023, doi: 10.2307/JIE.32.2.122.

[40] P. Rezaee Borj, K. Raja, and P. Bours, "Detecting Online Grooming By Simple Contrastive Chat Embeddings," in Proceedings of the 9th ACM International Workshop on Security and Privacy Analytics, 2023, pp. 57–65. doi: 10.1145/3579987.3586564.

[41] S. Craven, S. Brown, and E. Gilchrist, "Sexual grooming of children: Review of literature and theoretical considerations," Journal of Sexual Aggression, vol. 12, no. 3, pp. 287–299, 2006, doi: 10.1080/13552600601069414.

[42] P. Alonso-Ruido, I. Estévez, B. Regueiro, and C. Varela-Portela, "Victims of Child Grooming: An Evaluation in University Students," Societies, vol. 14, no. 1, p. 7, 2024, doi: 10.3390/soc14010007.

[43] A. Bull and T. Page, "Students' Accounts of Grooming and Boundary-Blurring Behaviours by Academic Staff in UK Higher Education," Gender and Education, vol. 33, no. 8, pp. 1057–1072, 2021, doi: 10.1080/09540253.2021.1884199.

[44] M. D. Cavelty, Cyber-security and threat politics: US efforts to secure the information age. Routledge, 2007.

[45] H. A. Kruger and W. D. Kearney, "A prototype for assessing information security awareness," Computers & Security, vol. 25, no. 4, pp. 289–296, 2006, doi: 10.1016/j.cose.2006.02.008.

Understanding Echo Chambers in Recommender Systems: A Systematic Review

Meriem HASSANI SAISSI[1, *], Nouhaila IDRISSI[2], Ahmed ZELLOU[3]
SPM, ENSIAS/Mohammed V University in Rabat, Rabat, Morocco[1, 3]
Department of Computer Science, Laboratory LAROSERI-Faculty of Sciences, El Jadida, Morocco[2]

***Abstract*—Echo chambers refer to the phenomenon in which individuals are consistently exposed to content that aligns with their existing viewpoints. Over time, this can narrow a user's perspective and make it harder to encounter different opinions. In this systematic literature review, we looked at studies published between 2019 and early 2025 and how they have approached this issue, from understanding the cause, and examining existing detection and mitigation strategies. We went through and organized the main findings, noting patterns in the algorithms used, the role of user behavior, and the influence of the data itself. Several works also suggest ways to introduce more variety into recommendations, aiming to break repetitive exposure. Our review confirms that echo chambers and filter bubbles do exist in recommender systems and that they raise concerns for diversity and fairness. Furthermore, we end by pointing to open questions and possible directions for future work, for both researchers and practitioners.**

***Keywords*—Echo chamber; recommender systems; filter bubbles; collaborative filtering; systematic literature review**

I. Introduction

Since their appearance in the 1990s, recommender systems (RS) have experienced huge evolution, evolving from basic filtering algorithms to complex models that influence user experiences on popular digital platforms. In order to provide recommendations, RS first used content-based filtering, which examined item descriptions. Collaborative filtering was introduced by GroupLens and became the standard method [1].

By introducing item-based collaborative filtering in the early 2000s, Amazon transformed the way suggestions were created by emphasizing item similarities [2], [3]. The use of deep learning in the 2010s was the next significant advancement, enabling recommender systems to recognize involved patterns in user behavior and content. Methods like wide & deep learning models and YouTube's neural networks [4] gained popularity. In order to provide contextual and sequential recommendations based on dynamic user interactions, recommender systems nowadays make use of sophisticated models like transformers [5] and reinforcement learning [6]. The discipline of recommender systems is concentrating on enhancing personalization while simultaneously addressing new issues with transparency, fairness, and diversity as a result of ongoing advancements in deep learning and hybrid models [7], [8].

When people are mostly exposed to information, viewpoints, or content that supports their preexisting preferences or ideas, this is referred to as an "echo chamber". This effect occurs when algorithms use user data to customize suggestions, emphasizing material related to the user's past interactions. Although this customization increases user engagement and pleasure, it also runs the risk of isolating consumers in uniform informational settings. Echo chambers are not a novel idea; they have been thoroughly examined in e-commerce [11], [12], social media [9], [10], and news and media [13]. This phenomenon was first examined by Sunstein, C. R., in the context of political discourse, emphasizing how homogenous societies can be produced and ideas polarized by selective information exposure [14].

According to research [15] on filter bubbles, echo chambers occur in the context of recommender systems when algorithms favor information that is similar to what consumers have already interacted with. The intention behind this selective exposure is to increase user pleasure and engagement; however, it unintentionally reduces the range of information that users can access. Promoting a more informed and well-rounded user base requires both understanding and actively mitigating the effects of echo chambers in recommender systems. The dynamics of algorithmic recommendation and its consequences for information variety have been studied recently in research like [12], [16]. The necessity of creating recommender systems that strike a balance between personalization and the incorporation of varied and possibly difficult content is highlighted by this research. In order to document recent developments and new trends in the industry, this analysis focuses on the last seven years. From an initial pool of 564 papers gathered from top scientific databases such as Scopus, IEEE Xplore, Google Scholar, SpringerLink, ACM Digital Library, and DBLP, 56 peer-reviewed studies published between 2019 and early 2025 were chosen for in-depth research.

In light of this, we provide an organized synthesis that is directed by important research topics, illuminating the primary strategies and techniques used in this field of study. Our contribution to this discussion is a Systematic Literature Review (SLR) conducted to explore strategies for reducing the impact of echo chambers in recommender systems and to identify their existence.

The remainder of this study is structured as follows: Section II provides essential background on recommender systems and the concept of echo chambers. Section III presents the related works on echo chambers in recommender systems and recent AI-driven approaches proposed to mitigate them. Section IV details the research methodology and the guiding research questions. Section V presents the main findings and

*Corresponding author.

offers a critical analysis of the selected studies, highlighting the presence of echo chambers in recommender systems and evaluating the effectiveness of different mitigation strategies. In Section VI, we discuss the implications of our findings. We explore the overall consequences of echo chambers on user behavior and information diversity, and we identify potential directions for future research to address limitations and challenges identified in the literature. Finally, Section VII concludes the study by summarizing the main insights and discussing potential directions for future research. We summarize the main contributions of our review and suggest approaches for future studies to further explore and mitigate the effects of echo chambers in recommender systems.

II. Background

In this section, a comprehensive overview is presented regarding the two primary areas of research that are directly relevant to this systematic review: recommender systems and the echo chamber problem.

A. Recommender Systems

1) Content-based filtering: This method generates recommendations by focusing on product features and user profiles. It analyzes the attributes of items a user has previously interacted with and suggests similar ones based on these characteristics. Aggarwal et al. provided a comprehensive analysis of content-based approaches and their applications, noting their effectiveness across various domains [17].

2) Collaborative Filtering (CF): One of the most popular techniques in recommender systems, which is predicated on the notion that customers with comparable tastes will appreciate comparable products. It looks at user activity, such as ratings and interactions, to provide personalized predictions. Collaborative filtering is commonly divided into two main types: user-based and item-based. In the user-based approach, the system identifies individuals with similar preferences to the target user and recommends items favored by those individuals. The basic idea is simple: if two people have liked similar things in the past, one is likely to like what the other liked. A prominent early contribution in this domain is the research conducted by Sarwar et al. [18], which applied user-based collaborative filtering to movie recommendations and evaluated its effectiveness in that context. Item-based collaborative filtering, on the other hand, focuses on the relationships between items rather than users. It recommends items that are similar to those the user has liked in the past. This approach calculates the similarity between items based on user interactions and suggests items that are most similar to the ones the user has previously rated highly. An influential paper by Koren et al. [19] demonstrates the effectiveness of item-based collaborative filtering, particularly in handling large-scale datasets and providing accurate recommendations. In addition to these traditional approaches, recent advancements have introduced techniques such as matrix factorization and neighborhood-based methods. Matrix factorization techniques, including Singular Value Decomposition (SVD), decompose the user-item interaction matrix into latent factors to discover hidden patterns and improve recommendations. The work by Koren et al. provides a comprehensive overview of matrix factorization techniques and their application to collaborative filtering [20]. Additionally, approaches that integrate collaborative filtering with other methods, such as hybrid systems, aim to address limitations like data sparsity and cold start problems, enhancing the overall recommendation accuracy [21].

3) Hybrid recommender systems: Combine the advantages of collaborative filtering and content-based filtering to improve recommendation performance. Hybrid systems aim to overcome the limitations of each method, such as data sparsity and limited diversity, by combining multiple strategies. In order to enhance recommendation quality, Liang et al. present a hybrid strategy that combines sentiment analysis with matrix factorization [22]; Burke [21] provides a comprehensive overview of various hybrid techniques and their applications. Hybrid methods combine the strengths of content-based and collaborative filtering, offering a more reliable and comprehensive recommendation experience.

Despite their advantages, these methods encounter substantial limitations, particularly in reinforcing echo chambers, which restrict user exposure to diverse content by continually highlighting existing preferences. In content-based filtering, echo chambers emerge when recommendations are primarily based on a user's past interactions, resulting in repetitive suggestions that gradually limit exposure to diverse perspectives [48]. The echo chamber effect is amplified by exclusive dependence on historical user and item features, which means that it rarely provides novel or diverse content [23]. Similarly, collaborative filtering perpetuates echo chambers by generating recommendations based on the preferences of similar users. As a result, users are repeatedly exposed to content that caters to the preferences of a small, similar group, resulting in a closed feedback loop. As a result, users encounter content that reinforces their biases, limiting exposure to different viewpoints [18].

Hybrid models may continue to highlight user preferences over diversity in recommendations without deliberate strategies [21]. This would reinforce narrow content patterns and limit diversity. In conclusion, despite their benefits, these strategies underscore the ongoing issue of echo chambers in recommender systems and the need for new approaches that actively promote diverse content and diversify user exposure.

B. Echo Chamber Problem

An echo chamber, also known as filter bubble, is a phenomenon commonly observed in recommender systems, where users with similar beliefs mainly interact with each other, reinforcing shared viewpoints and potentially intensifying those opinions. This concept is likened to an echo chamber where homogeneous views are repeatedly echoed within the group, which can sometimes lead to more extreme beliefs [24]. Echo chambers are feared to exacerbate political polarization and public opinion division [25]. Echo chambers are typically identified through their structural characteristics, such as the density of connections and network homophily, as well as the ideological uniformity of their members [28], [33]. Research

utilizing digital trace data has examined these features to understand their role in online political polarization and the spread of misinformation [42], [43], [44]. Although some studies suggest that the existence of echo chambers may be overstated [46], [47], evidence generally shows that users tend to form communities with closely aligned views [45].

III. Related Works

The concept of "echo chambers" in recommender systems has evolved from earlier discussions in communication theory, which examines how individuals seek information that confirms their pre-existing views. Cass Sunstein was one of the first to introduce the term "echo chamber" in his work *Republic.com*, where he expressed concern that online algorithms could create isolated communities of like-minded individuals by limiting exposure to diverse perspectives [14]. This concern grew with the increasing influence of algorithms, as Eli Pariser's in "*The Filter Bubble*" [15], the concept is taken further, showing how personalization in recommender systems, especially on social media, can shape the flow of information to fit individual preferences. Over time, this selective exposure can limit the range of viewpoints users encounter, leading to ideological isolation and creating echo chambers [41].

Additionally, the role of echo chambers in the spread of misinformation became particularly prominent during the COVID-19 pandemic, as demonstrated by Cinelli et al., who analyzed how social media platforms amplified false information through echo chambers during the crisis [28].

From 2023 to 2025, recent studies have deepened our understanding of how recommender systems can shape and reinforce echo chambers, while others have explored ways to counter these effects. Gao et al. [37] showed that short-video platforms such as Douyin and TikTok often trap users in repetitive content loops created by engagement algorithms. Similarly, Duskin and Schäfer [67] found that Twitter's Who-to-Follow feature strengthens social homophily, leading to more polarized online communities. Sharma et al. [71] warned that conversational search tools powered by large language models could unintentionally create new forms of selective exposure, while Iwanaga et al. [73] demonstrated that friend-recommendation algorithms can reproduce echo-chamber structures in social networks. Collectively, these studies highlight that echo-chamber effects emerge from continuous user algorithm feedback loops and that mitigating them requires AI-driven designs that promote diversity and balanced information exposure.

While significant progress has been made in understanding echo chambers and filter bubbles in digital environments, existing systematic literature reviews (SLRs) reveal several limitations. Table I presents a summary of key limitations in notable SLRs and the added value of this study.

This SLR provides significant contributions by bridging the gaps left by previous studies. Unlike earlier reviews, it examines a broad timeline, covering research from 2019 to 2025, ensuring inclusion of the most recent advancements. My work integrates new technologies, highlighting their potential in addressing echo chambers effectively.

TABLE I. Identified Limitations in Existing Echo Chamber Studies

Ref.	Limitation
[52]	Does not address recommender systems and provides limited insights into mitigation strategies.
[53]	This review has several limitations. Many studies do not explicitly address the term "echo chamber" and rely on older approaches that may not reflect recent advances. In addition, the review dates back to 2022, while the field has since experienced rapid, exponential growth.

IV. Research Methodology

A. The Need of an SLR

A systematic literature review (SLR) is a structured and comprehensive way of bringing together what has already been studied on a given topic. In academic research, it helps identify and interpret all relevant work connected to a particular question, field of interest. However, past SLRs on recommender systems in echo chambers have left significant limitations. They have overlooked key studies and recent developments that are essential for a complete understanding, and many have failed to include all the necessary literature. Important areas that were often missing include the use of reinforcement learning and deep learning techniques to address echo chambers.

New research directions are especially important for a fuller picture. These include generative AI to counter polarization and adopting reinforcement learning to encourage more unexpected content. Another emerging strategy is to maximize semantic variety in recommendations as a way to limit echo chamber effects. Unfortunately, many earlier SLRs relied on outdated and incomplete methods, leading to biased conclusions. This review aims to close those limitations by using a variety of sources and adopting more recent, comprehensive research methods.

B. Stages of the SLR

In the field of software engineering, Kitchenham's [30] guidelines offer a clear and practical framework for conducting systematic literature reviews. This approach allows researchers to combine evidence in a way that is both structured and repeatable. As shown in Fig. 1, the process involves several key steps: defining the research questions, preparing a review protocol, performing an extensive literature search, applying predefined inclusion and exclusion criteria to select studies, assessing the quality of those studies, extracting the relevant data, and finally, analyzing and synthesizing the findings.

The systematic literature review (SLR) process is typically organized into three main stages: planning, conducting, and reporting. In the planning stage, the need for the review is established, followed by defining the research questions, keywords, and search terms. Researchers then decide on the information sources, develop study selection criteria, and design data extraction forms. In the conducting stage, the relevant studies are chosen, their quality is evaluated, and the necessary data is extracted and synthesized to draw meaningful conclusions. Finally, in the reporting stage, the findings are analyzed, compiled into a synthesis report, and verified to ensure both accuracy and reliability.

C. Research Objective

The questions and objectives that our SLR will address, as shown in Table II, will focus on understanding the echo chamber phenomenon in recommender systems, evaluating existing mitigation strategies, examining new trends and techniques, and identifying unexplored areas in the literature to guide future research efforts. The objective is to provide a comprehensive and up-to-date understanding of the topic.

We used two distinct search strategies to gather as much evidence as possible when conducting the SLR: ("Echo Chamber" OR "Filter Bubble") AND ("Recommender System" OR "Recommendation System"). Scopus, IEEE Xplore, Google Scholar, SpringerLink, the ACM Digital Library, and DBLP were the first six electronic data sources that we searched automatically. We were able to gather a list of potential candidates thanks to this strategy. These databases were chosen because they contain all high-quality conference proceedings and journals related to recommender systems. Additionally, following best practices for systematic literature reviews (SLRs), we used the snowballing search strategy. This required looking over the bibliographies of relevant papers and utilizing them as a foundation to locate additional relevant literature.

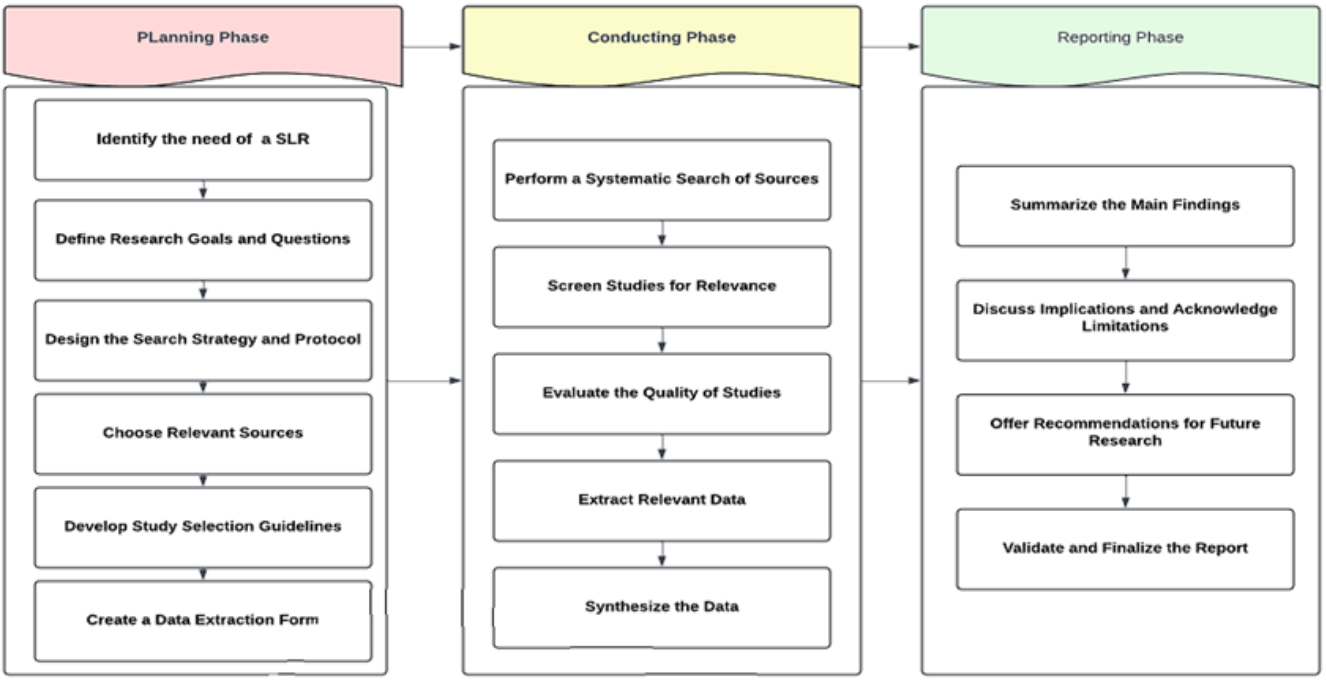


Fig. 1. Key elements of every phase in the SLR.

TABLE II. RESEARCH QUESTIONS

Research Questions	Motivation
RQ1: How echo chamber can be defined and detected in different recommendation approaches?	To comprehend how various recommendation systems create echo chambers. Different types: "collaborative filtering", "content-based", and "hybrid" may reinforce user biases in distinct ways. We can find better ways to reduce their impact by defining these.
RQ2: What approaches are implemented to prevent echo chambers in recommender systems, and what techniques are utilized to achieve this	To look into ways to prevent recommender systems from becoming echo chambers. The diversity of recommendations can be increased and users will not be exposed to content that is either biased or repetitive if they are aware of the techniques that work best. We can make recommendations more diverse and improve the user experience by identifying these methods.
RQ3: Have there been any evaluations? If so, what metrics were employed to assess the presence of echo chambers?	To determine whether any evaluations of echo chambers have been carried out. The purpose of this RQ is to identify the metrics used to evaluate the presence of echo chambers.
RQ4: Which datasets were used for evaluating echo chambers?	To identify the datasets commonly used in evaluating echo chambers within recommender systems. Understanding the datasets can provide insights into their suitability and limitations.
RQ5: What are the most promising directions for further research?	To examine the limitations of current solutions and identify future research directions regarding echo chamber and filter bubbles in recommender systems.

D. Selection Process and Eligibility Criteria

We began our selection process by reviewing the studies' titles, publication years, keywords, and abstracts. Then, we applied the criteria outlined in Table III to decide whether each study should be retained for further analysis. Any document that met at least one of the exclusion conditions was discarded.

TABLE III. DETERMINE THE CRITERIA FOR INCLUDING AND EXCLUDING STUDIES

Inclusion criteria	Exclusion criteria
The relevance of each paper was assessed according to its contribution to the five research questions (RQ1 to RQ5).	Papers outside the scope, as they do not focus on RSs
Papers from journals or conferences	Papers that address recommender system but not echo chamber
Papers published between 2019 and 2025	Duplicate papers
Language: written in English	Non-Academic Sources

TABLE IV. QUALITY QUESTIONS

Inclusion criteria	Exclusion criteria	Weight
Are the study's key contributions and motivations stated clearly and precisely?	Yes/partly/no (1/0.5/0)	1
Is the proposed solution clearly explained and feasible for implementation?	Yes/partly/no (1/0.5/0)	1
Did the study clearly define and thoroughly discuss the proposed solution?	- The study provided a high-level overview of the proposed system but lacked in-depth technical details (0.5). - The study offered a clear description of the methodology but limited explanation of its real-world application (1). - The proposed solution is mentioned, but key components are not fully developed or explained (0.5). -The proposed solution is not mentioned or referenced in the study (0).	2
Does the study compare the proposed solution with other existing works?	Yes/partly/no (1/0.5/0)	1
Does the study provide empirical evaluation and present results?	-The study proposed a solution but offered no implementation or evaluation (0). -The study provided an implemented solution but failed to address proper evaluation (0.5). -The study offered a fully implemented and well-evaluated solution. (1).	1,5
Are the study's findings clearly stated?	Yes/partly/no (1/0.5/0)	1

E. Quality Assessment (QA)

To maintain the credibility of our systematic literature review, we assessed the quality of each selected study using a set of predefined questions outlined in Table IV. For each question, responses were scored as 0 (no), 0.5 (partially), or 1 (yes), which allowed us to evaluate how well each criteria was met. Since some questions carried more importance than others, we applied specific weightings accordingly. The overall quality score for each study was then calculated using Eq. (1), combining both the assigned scores and their respective weights.

$$\text{Score} = \sum_{j=1}^{6} \frac{v_j * w_j}{7{,}5} \times 100 \quad (1)$$

where, w_j the weight assigned to each criterion, reflecting its importance and the value assigned to each quality criterion. Based on our quality assessment, we retained only the studies that scored at or above the average. This process resulted in a final selection of 56 papers for inclusion in our review.

F. Data Extraction Strategy and Synthesis

The relevant information from the final selected papers was documented using the data extraction form shown in Table V. This was done to address the research questions (RQs) outlined in Table IV and to accurately capture the data gathered during the study.

TABLE V. DATA EXTRACTION FORM

Data extraction form
ID, Title, Authors Publication type (Journal/Conference), Publication Year Domain or application feld, Contribution: Findings and Outcomes The study quality assessment **RQ1**.Definition in different recommendation type **RQ2**. Approaches and Techniques used **RQ3**. Evaluation measures and results **RQ4**. Datasets **RQ5**. Limitations & Future work

V. RESULTS

A. Overview of the Selected Studies

After completing the first step of the study selection process, as shown in Fig. 4, a total of 752 publications were retrieved from the Electronic Data Sources (EDS) searches using the search string. The titles of the articles were then checked for duplicates, and 564 candidates remained for further refinement based on the previously explained inclusion and exclusion criteria. The number of research papers to be reviewed in the systematic review was reduced from 564 to 234 after applying the exclusion criteria, and only 56 of these were found to be relevant. An additional six articles were identified by reviewing the reference lists of the relevant papers. Finally, the six quality assessment criteria listed in Table III were applied to ensure that the included findings would contribute valuable insights to the systematic literature review (SLR). Eight papers, including [75] [87] [93] [98] [100] [102] [104] [108], that scored 53.68 or lower, were discarded.

B. Classification of the Selected Studies

In this phase of the research, we investigate the distribution of the selected studies by years of publication, data sources and publication venues.

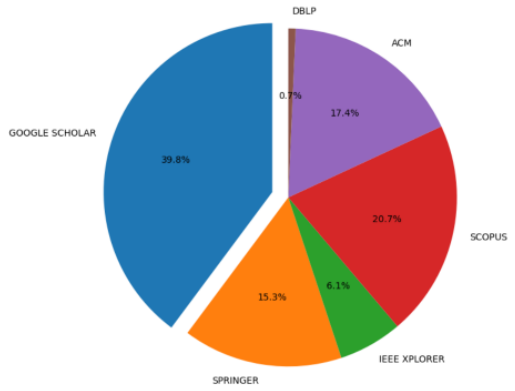


Fig. 2. Distribution by data sources.

Fig. 2 shows the proportion of selected research papers from various data sources. Google Scholar accounts for the largest share, providing 39.8% of the relevant papers, followed by Scopus with 20.7%, and ACM with 17.4%. Springer contributes 15.3%, while IEEE Xplorer and DBLP provide 6.1% and 0.7%, respectively. This distribution emphasizes the importance of Google Scholar and Scopus as major sources for gathering relevant research in this domain.

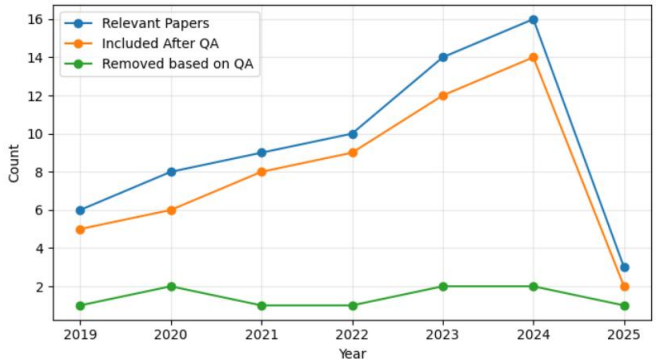


Fig. 3. Distribution of relevant papers before and after QA (2019-2025).

Fig. 3 illustrates the distribution of relevant papers from 2019 to early 2025, highlighting both the papers included after the quality assessment (QA) and those removed based on the QA. The total number of relevant papers increased from 7 in 2019 to a peak of 16 in 2024, showing growing interest in the field. Papers included after QA also rose from 5 in 2019 to 14 in 2024. The number of papers removed during QA fluctuated between 1 and 2 over the same period. In early 2025, both relevant papers and those included after QA dropped to 3 and 2, respectively, while papers removed during QA stood at 2.

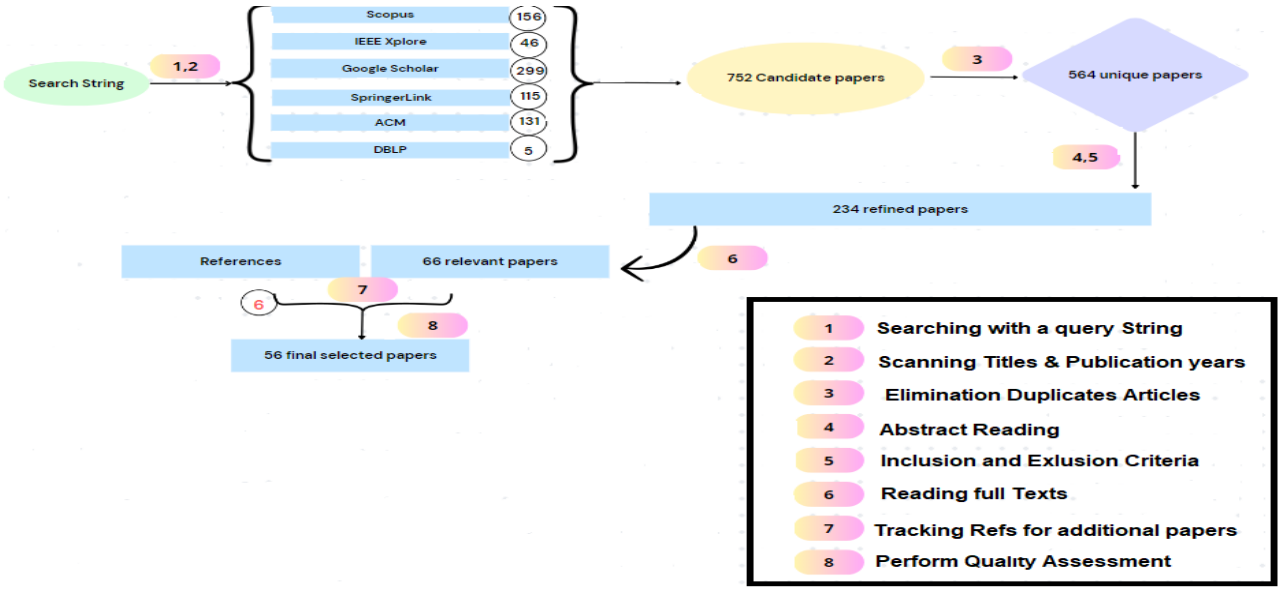


Fig. 4. Summary of the final paper selection process.

C. Data Extraction Results

In the following sections, we will present the findings of this systematic review and analyze them to address the research questions outlined in Section IV.

1) Summary of the RQ1: Echo chambers are a significant concern in various types of recommender systems, each manifesting in distinct ways depending on the underlying algorithm. In collaborative filtering systems, echo chambers arise when users are consistently exposed to content that mirrors their past interactions and the preferences of similar users. This feedback loop intensifies the reinforcement of existing beliefs, leading to a narrowing of perspectives and a reduction in content diversity. The concept of echo chambers in collaborative filtering has been extensively examined in recent literature, highlighting both the mechanisms behind their formation and the implications for user experience. Yingqiang Ge et al. [58] describe an echo chamber as a phenomenon in which users' interests are reinforced through repeated exposure to similar items or categories. This process closely aligns with

the principles of collaborative filtering, where user interactions and preferences are analyzed to generate recommendations based on the behavior of similar users. In addition, Ruben Interian et al. highlight the feedback loop embedded in collaborative filtering systems: as user preferences are continually reinforced, individuals are repeatedly exposed to similar content. Over time, this cycle restricts the diversity of perspectives and encourages to echo chamber [31]. Similarly, Sami Khenissi et al. [88] explain that echo chambers often emerge from the algorithmic tendency to recommend content that aligns with a user's previous behavior, gradually narrowing the range of information available to them. Alireza Gharahighehi et al. focus on session-based recommender systems, underscoring how the emphasis on predictive accuracy in collaborative filtering can inadvertently lead to the creation of echo chambers by exposing users to similar content and viewpoints over time [49]. Mert Can Cakmak et al. [107] Defines echo chambers as situations where algorithms repeatedly expose users to similar viewpoints, reinforcing pre-existing preferences. It also discusses detection via analyzing homogeneity in user–item interactions and network-based similarity. Lastly, Vancompernolle Vromman and Fouss [96] provide an experimental analysis, demonstrating that collaborative filtering algorithms naturally reduce content diversity, thereby reinforcing existing user preferences and contributing to the echo chamber effect. A common theme emerges from all of these studies: while collaborative filtering improves user satisfaction through relevant recommendations, it also runs the risk of creating closed content loops that discourage engagement with a wider range of perspectives [15]. Similarly, in content-based filtering systems, echo chambers develop as the algorithm prioritizes content similar to what the user has previously consumed. Antoine Vandelville et al. [40] define an echo chamber in this context as a situation where users are predominantly exposed to content that aligns with their existing beliefs and preferences, leading to a reinforcement of those views. This occurs because content-based algorithms curate and recommend items based on the user's past interactions, often resulting in a narrow range of viewpoints being presented. As users engage with content that reflects their interests, the recommendation system continues to suggest similar content, thereby creating a feedback loop. This can limit exposure to diverse perspectives and opposing ideas, fostering an environment where users remain insulated within their own ideological bubbles. Consequently, echo chambers in content-based systems can hinder open discourse and contribute to the polarization of opinions.

Hybrid recommender systems, which combine elements of both collaborative and content-based filtering, attempt to mitigate some of these issues by balancing accuracy with diversity. However, they are not immune to the creation of echo chambers, particularly if the system overemphasizes content that aligns with the user's established preferences. The challenge with hybrid systems lies in effectively balancing the competing demands of personalization and diversity, which, if not carefully managed, can still lead to the reinforcement of homogeneous content [49, 96]. Given these critiques, it is evident that each type of recommender system has inherent strengths and weaknesses concerning the formation of echo chambers. To mitigate the risks associated with echo chambers, strategies such as introducing diversity and serendipity in recommendations, employing algorithms that balance user exploration with relevance, and ensuring a mix of content that includes both popular and lesser-known items are essential. Additionally, applying fairness metrics to prevent the disproportionate favoring of specific content types can help create a more balanced recommendation environment [12, 62].

The next summary of RQ2 will delve deeper into how these mitigation strategies can be implemented across different recommender systems to reduce the impact of echo chambers, ensuring a more inclusive and diverse content experience for users.

2) Summary of the RQ2: After analyzing how researchers define the echo chamber, we now focus on the review of proposed of proposed approaches in collaborative filtering and other as well as the identification of the techniques used to avoid echo chamber. In the research area of collaborative recommendation and others types, several techniques have been disclosed for avoiding echo chamber. Table VI classifies studies according to techniques used for overcoming echo chamber issues in both collaborative approaches and other.

Collaborative Filtering based approaches leverage user-item interactions to generate recommendations, with several techniques focusing on enhancing diversity to prevent echo chambers. One such technique is the Targeted Diversification VAE-based Collaborative Filtering (TD-VAE-CF), which integrates Variational Autoencoders to inject targeted diversification into the recommendation process [60]. Reinforcement Learning based methods are also applied within CF frameworks, where adaptive learning helps to diversify recommendations over time by considering long-term user satisfaction rather than short-term interests [36], [38], [61]. Another noteworthy CF-based method is the Cross-domain.

Recommendation Model with Adaptive Diversity Regularization, which extends the concept of diversity across different domains, thereby broadening the scope of recommendations beyond a single domain [29]. Additionally, Novel serendipity-oriented RS, Generating Self-Serendipity Preference (GS^2-RS) is designed to introduce serendipitous recommendations, making the user's experience more varied and less predictable [65].

On the other hand, other types of Recommender Systems incorporate approaches that are not strictly based on collaborative filtering but are equally crucial in avoiding echo chambers. The Gini Coefficient Analysis is utilized to measure inequality in the distribution of recommendations, thus ensuring a fairer spread of content exposure across different users [81]. The Community Aware Model [31], [66], [81], [90] focuses on the social aspects of recommendations by accounting for the community structure, which inherently promotes a diversity of perspectives. Techniques like Determinantal Point Process (DPP) provide a probabilistic

framework that naturally favors diverse outcomes in recommendation generation [32], [88]. Furthermore, Contrarian Recommendations deliberately suggest items contrary to the user's typical preferences, thereby breaking the cycle of repetitive content exposure [5].

Other advanced approaches include Maximizing Semantic Volume, which aligns with the content-based approach by expanding the semantic range of recommendations. This technique ensures that users are exposed to a broader variety of topics, mitigating echo chambers and enhancing diversity within their content consumption [20, 26]. Similarly, Diversification Approaches for Session-Based Recommenders are often integrated into hybrid approaches, combining elements of content-based and re-ranking strategies to dynamically adjust recommendations during a single user session, maintaining engagement and novelty [20, 26, 49]. Allostatic Regulator introduces an adaptive framework under the content-based approach, balancing content exposure dynamically to ensure a healthy mix of familiar and novel recommendations, further addressing echo chamber concerns [39, 54]. Within the knowledge-based approach, techniques like GeNeG (Generalization via Sentiment and Stance Detection) leverage sentiment and stance analysis to introduce diverse perspectives into recommendations. This method promotes balanced content exposure, directly tackling polarization by presenting varied viewpoints based on explicit content analysis [84]. Similarly, the Point-of-View Diversification Technique falls under the knowledge-based approach, intentionally incorporating content from diverse perspectives to promote diversity and user awareness [51].Content-based approaches also include techniques such as Linguistic Diversity, which identifies and incorporates semantically diverse content into recommendations, and Clustering Models using TF-IDF, which focus on grouping content features to ensure varied exposure [4, 20]. These approaches highlight item-specific attributes to balance user preferences with novel suggestions. Finally, advanced hybrid approaches like the Re-ranking are instrumental in reordering recommendations to maximize diversity and fairness while maintaining relevance [70, 94, 107, 107]. Hybrid approaches effectively leverage advanced techniques such as Multi-Agent Systems with Large Language Models (LLMs) [77] and Counterfactual Interactive Recommender Systems (CIRS) [39] to address echo chambers and enhance user experience through innovative interaction and modeling techniques. These diverse methods underscore the multifaceted strategies being employed to enhance recommendation diversity, ensuring that users are introduced to a more diverse set of content and viewpoints, thus effectively mitigating the risk of echo chambers in recommender systems.

3) Summary of the RQ3: Having discussed the different approaches for mitigate echo chamber, we now proceed to the investigation of how these approaches can be validated. Mathematical metrics play a crucial role in quantifying and evaluating the impact of recommendations on user behavior and content diversity.

Diversity metrics, such as ILD@k and the Gini-Simpson Index, are widely referenced in [80], [85], [99], and [49]. These metrics assess how varied the recommended content is. For ILD@k measures the average dissimilarity of items within a list, while the Gini-Simpson Index evaluates the probability that two items belong to different categories. High diversity metrics indicate that the system is exposing users to a broader range of content, reducing the likelihood of echo chambers by encouraging diverse content consumption.Coverage, discussed in [27], [32], [72], and [80], measures the proportion of the item space that is recommended to users. High coverage implies that a wide range of content is being suggested, which helps prevent echo chambers by ensuring that users are exposed to a variety of content rather than just a narrow subset.

The Precision and Recall metrics, as referenced in [27], [32], [38], [85], and [99], are commonly used to evaluate the accuracy of recommender systems. Precision measures the proportion of recommended items that are relevant, while Recall assesses the system's ability to retrieve all relevant items. While these metrics focus on relevance, there is a need to balance them with diversity metrics to avoid reinforcing echo chambers. High precision and recall alone can sometimes contribute to echo chambers if the recommendations consistently reflect known user preferences without introducing new or diverse content.

NDCG (Normalized Discounted Cumulative Gain), as mentioned in [27], [32], [38], and [60] and defined in Eq. (2) and Eq. (3), evaluates the ranking quality of recommended items, taking into account both relevance and the position of items in the recommendation list. This metric is important for ensuring that while the recommendations are accurate, they also maintain a balance with diversity to prevent the dominance of a narrow set of content. Formally, it is defined as:

$$DCG = \sum_{i=1}^{K} \frac{rel_i}{\log_2(i+1)} \quad (2)$$

$$NDCG = \frac{DCG}{IDCG} \quad (3)$$

where, rel_i is the graded relevance score of the i^{th} item and $IDCG$ is ideal discounted cumulative gain [60].

MAE (The Mean Absolute Error), defined in Eq. (4), computes the deviation between the actual and predicted ratings and RMSE (Root Mean Square Error), defined in Eq. (5), discussed in [82], measures the accuracy of predictions by calculating the square root of the average squared differences between predicted and actual values. While RMSE is primarily focused on accuracy, it is important to consider that overly optimizing for RMSE without attention to diversity can contribute to echo chambers by continuously reinforcing existing preferences.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|p_{i-r_i}\right| \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - r_i)^2} \quad (5)$$

The Isolation Index (Iso-Index), highlighted in [27] and [72], measures the segregation of users into isolated groups receiving similar recommendations. A high Iso-Index suggests that users are isolated within echo chambers, as they are only exposed to content that aligns closely with their existing preferences.

TABLE VI. TECHNIQUE USED TO MITIGATE ECHO CHAMBERS IN RS

Refs.	Technique used	Type of approach			
		Collaborative	*Hybrid*	*Modeling*	*other*
[5]	Contrarian recommendations				Novelty Based
[20],[40]	Maximizing semantic volume				Content Based
[27]	New Prototype User Controllable RS			✖	
[29]	Cross-domain recommendation model with adaptive diversity regularization	✖			
[32],[88]	Determinantal Point Process (DPP)			✖	
[32]	Contrastive Co-training	✖	✖		
[34]	Clustering Distance-based Method				
[35]	IPS-based sequence recommendation system (IPS-seqRS)	✖			
[36], [43],[99]	Graph neural network: FRediECH				Graph Based
[36], [38], [61],[106]	Reinforcement Learning based method	✖			
[39]	CIRS, a counterfactual interactive recommender system		✖		
[49]	Candidate Item Selection Manipulation Neighbor Selection Manipulation		✖		Content Based
[50]	Multi-Perspective Search Paradigm	✖	✖		Knowledge-Based
[51]	A point-of-view diversification technique	✖			Knowledge-Based
[54]	Allostatic regulator			✖	
[56]	Modified Random Walk Exploration (RWE)				Graph-Based
[60]	Targeted Diversification VAE-based Collaborative Filtering (TD-VAE-CF)	✖			
[65]	Novel serendipity-oriented RS called GS^2-RS	✖			
[66]	RG REC by integrating Generative Artificial Intelligence (GAI)				Generative AI-Based
[70]	Diversity nudges		✖		
[34],[70],[94], [107]	Re-ranking		✖		
[77]	Multi-Agent Systems and Large Language Models (LLMs)				Hybrid/Generative AI-Based
[78]	SSLE framework		✖		
[81]	Gini Coefficient Analysis		✖		Evaluation Metric
[83]	Linguistic Diversity				Content-Based Diversity
[84]	GeNeG (Generalization via Sentiment and Stance Detection)				Knowledge-Based
[86]	A pseudo global social network				Community Detection
[95]	Inverse Propensity Weighting (IPW)	✖			
[97]	New metric based on the homogenization				Evaluation Metric
[103],[55]	Agent-Based Modeling (ABM)				Simulation-Based
[105]	Clustering based model : TF – IDF algorithm (ICDM)				Content Based

Clustering Coefficient, as discussed in [99], measures the degree to which nodes in a network tend to cluster together. High clustering indicates that users are part of tightly-knit communities, which can lead to echo chambers where users are only exposed to content within their group, reinforcing existing viewpoints.

DIS-EUC is mentioned in [27], and defined in Eq. (6). This metric measure the distance between the recommendations of different user groups. Larger distances suggest stronger group segregation and more severe echo chamber effects.

- $d_u \in R^M$: The distribution over item categories in the recommendations for user u.

- $\bar{g}_u \in R^M$:: The average distribution over item categories for the original group $\bar{x}$
- $\hat{g}_u \in R^M$: The average distribution over item categories for the target group $\hat{x}$

$$DIS - EUC = dis(d_u, \hat{g}_u) - (d_u, \bar{g}_u) \quad (6)$$

where, $dis(.)$ represents the Euclidean distance.

Finally, NCI (Normalized Clustering Index) and RWC (Relative Weight of Communities), both mentioned in [36], quantify how users are clustered based on similar opinions and the extent to which users are grouped into distinct communities. High values for these metrics indicate strong echo chambers, where users are more likely to interact with like-minded individuals and less likely to encounter diverse perspectives.

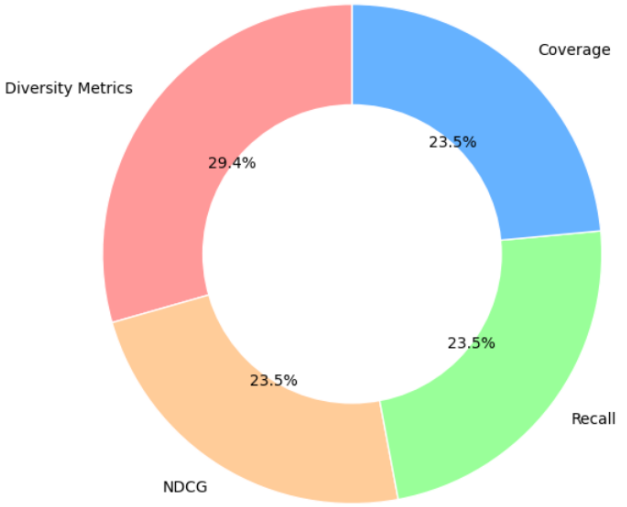


Fig. 5. Proportion of metrics used.

Fig. 5 provides a visual representation of the distribution of metrics used in evaluating recommender systems, particularly in the context of echo chambers. The chart shows that Diversity metrics are the most commonly used, diversity metrics. This distribution suggests a balanced focus on various aspects of recommendation quality, with a particular are also significant but are used slightly less frequently than making up 29.4% of the total. NDCG, Recall, and Coverage metrics each account for 23.5%, indicating that these metrics emphasis on ensuring diversity to combat echo chambers.

4) Summary of the RQ4: The choice of dataset for evaluating echo chamber in recommender systems depends on several factors, including the specific research goals, the characteristics of the recommender system being evaluated, and the availability of suitable data.

Table VII highlights the usage of various datasets in the study of echo chambers in recommender systems. The analysis shows that the MovieLens and Twitter datasets are the most frequently used in research aimed at mitigating echo chambers.

The MovieLens dataset is commonly utilized due to its rich data on user interactions with movies, making it a valuable resource for understanding how recommendation systems can influence user preferences and potentially create echo chambers. Similarly, the Twitter dataset is popular in research because it provides extensive information on social interaction, which is crucial for studying the dynamics of echo chambers in social media environments.

These two datasets play a dominant role in the field, highlighting their central importance in current efforts to investigate the challenges associated with echo chambers in recommender systems. The frequent use of these datasets underscores their importance in developing strategies to ensure that recommendations remain diverse and prevent the reinforcement of narrow user preferences.

Fig. 6 provides a visual representation of the distribution of datasets across different domains in the study of echo chambers in recommender systems. The key observations from the figure are:

Movies represent the most dominant domain, with the largest number of datasets. This suggests that a substantial portion of research on echo chambers focuses on movie recommendation systems, likely due to the widespread use of platforms such as Netflix and MovieLens. The prominence of these systems makes them a key context for examining how echo chambers develop and how they may influence user preferences and viewing habits.

Social media emerges as the second most prominent domain. Given the influence of social platforms on public discourse and user behavior, it is unsurprising that they remain an important focus in studies addressing the dynamics of echo chambers.

E-commerce and News domains each have a moderate representation. This highlights that echo chambers could affect consumer behavior in online shopping platforms and the information people consume from news sources.

Videos, Books, and Music have a smaller but still notable presence in the research. These domains show that echo chamber effects are being studied across various types of content, not just text-based or social media platforms.

Others represents a collection of miscellaneous domains that are less represented individually but still relevant to the overall study of echo chambers.

In summary, Fig. 6 highlights that research on echo chambers in recommender systems is focused on movies and social media with significant attention also given to e-commerce and news platforms. This suggests an interest in understanding how recommendation algorithms influence user behavior across a variety of digital content and platforms.

5) Summary of the RQ5: The last RQ of this SLR addresses future directions on echo chambers in recommender systems. The identified directions suggest various paths to improve how recommender systems function, focusing on mitigating the effects of echo chambers:

TABLE VII. THE DATASETS

Ref.	Dataset	Size of Dataset
[11],[27],[31], [34],[54],[63],[65],],[82],[89],[92], [106],[107]	Movie Lens	6,040 users, 3,702 movies, and 1 million movie ratings [31]: 100,000 ratings of 1682 movies [54] : 6,000 user ratings on 4,000 movies
[26] [40], [62] [81], [85],[90] [99],[107]	Twitter Dataset	[81]: 2.2 million nodes (users), 325.5 million edges (connections), and 3 billion tweets [26]: 2.2 million nodes (users), 325.5 million edges (connections), and 3 billion tweets [40] : 2,414,584 tweets and 7,763,931 retweets from 22,853 Twitter profiles [62]: approximately 217 million tweets
[27]	DIGIX-Video	N/A
	Amazonn-book	N/A
[29]	Douban	N/A
[32]	Amazon Music	Number of Users: 5,541 Number of Items: 3,568 Number of Interactions: 64,706 Number of Categories: 60
	Amazon Sport	Number of Users: 35,598 Number of Items: 18,357 Number of Interactions: 296,337 Number of Categories: 149
[38]	Gowalla	950,327 friend relationships among 196591 users
[38], [60]	Yelp2018	4.7 million user reviews, more than 150,000 merchant profiles, 200,000 images, and 12 metropolitan areas
[39]	Kuai Rec	N/A
[49]	Routana	Number of Sessions: 324,505 Number of Interactions: 5,529,007 Number of Items: 29,641
	Kwestie	Number of Sessions: 297,342 Number of Interactions: 922,680 Number of Items: 22,465
	Globo	Number of Sessions: 1,048,389 Number of Interactions: 2,986,226 Number of Items: 45,559
	Adressa	Number of Sessions: 5,607,303 Number of Interactions: 8,894,565 Number of Items: 27,805
[51]	The Brazilian presidential election	True news : G1: 905 news articles R7: 959 news articles Fake News: LUPA Agency: 78 news articles
[58]	Alibaba Taobao	Accessing logs from 86,192 users over a five-month period, spanning from January 1, 2019, to May 31, 2019.
[59]	-	40,000 news articles and a set of 500 users
[60]	Reddit	N/A
[64]	subset of the TREC Washington Post corpus	728,626 news articles and blog posts from January 2012 through December 2020
[72]	ReDial	N/A
[91]	TG-REDIAL	N/A
[92]	Yelp Challenge	N/A
[97]	Own Dataset	317 users registered in the system, and 221 interacted with the news recommendations

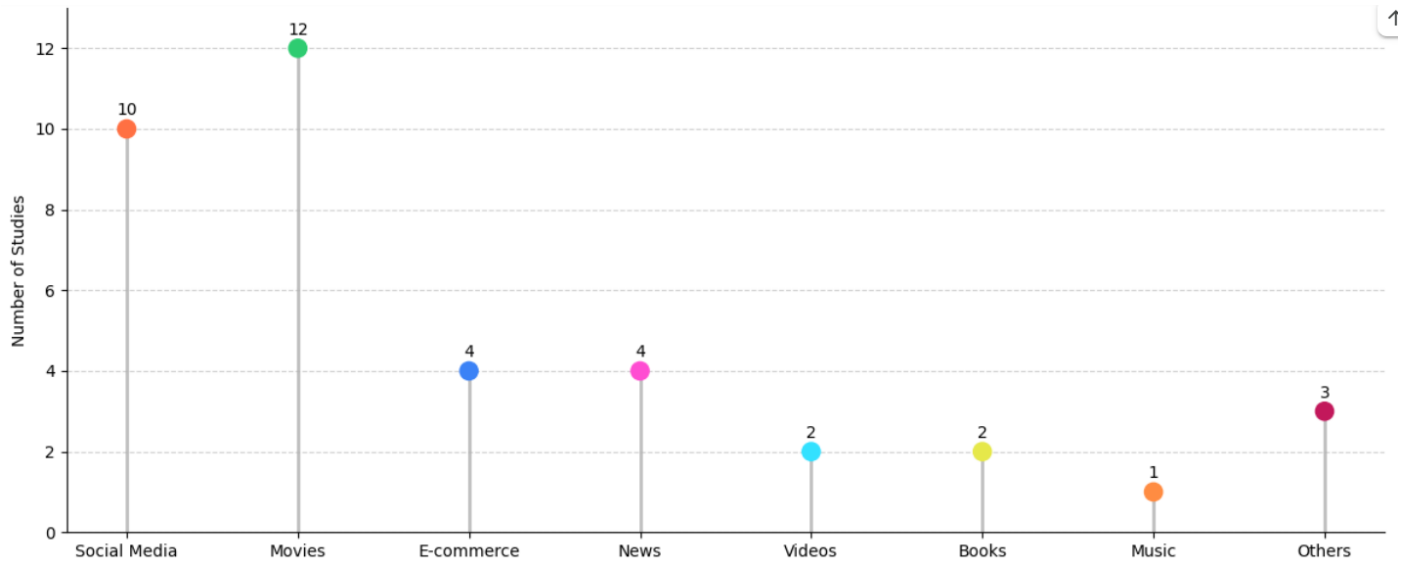


Fig. 6. Count of datasets by domain.

Longitudinal Studies (25%): Conducting long-term studies to track changes in user behavior over time is crucial. Such studies will help understand how echo chambers evolve and the long-term effects of potential interventions [26] [31] [68] [70] [78] [79] [89] [97].

Algorithmic Interventions (15%): Further development and testing of algorithms are necessary to promote diverse viewpoints. These new algorithms will be essential in combating echo chambers by exposing users to a wider range of content [68] [78] [79].

Empirical Validation (5%): A significant direction for future research is the need for more empirical studies that use real-world data from social media platforms to validate computational models. Comparing simulation results with actual user behavior will help determine the real-world impact of echo chambers [55] [69] [79].

Cross-Cutting Content Promotion (5%): Researchers should focus on developing strategies to promote content that bridges different perspectives. Encouraging dialogue between opposing viewpoints is essential for reducing polarization and creating a more balanced user experience [31] [89] [97].

Ethical and Interdisciplinary Approaches (5%): Future research should investigate ethical algorithm design and collaboration across disciplines for a holistic perspective. [26] [76] [78].

User Behavior and Engagement (20%): More research is needed on how users react to diverse content and opposing viewpoints. Understanding user engagement patterns will guide the design of recommender systems that foster more balanced and meaningful interactions [31] [49] [59] [79] [91] [103].

Advanced Technologies and Concepts (15%): Applying AI techniques, like generative AI and deep learning, to improve recommendations [5] [58] [74] [77] [78] [85].

User-Centric Approaches (10%): Designing systems that are tailored to individual users' preferences, behaviors, and feedback will be crucial for creating more personalized and diverse recommendation experiences [26],[78],[89].

An important future direction is the integration of generative AI to enhance content diversity and balance in recommendations. By leveraging generative AI, systems could produce content that represents multiple perspectives or simulate discussions reflecting diverse opinions, fostering a more inclusive and dynamic user experience [74],[101].

VI. Discussion

After following a specific series of detailed steps to collect the most relevant evidence, we now analyze and critique the findings. In addition, Section VI B outlines the limitations of this SLR. A summary of the results for each research question (RQ) is visually presented in the schematic model shown in Fig. 7.

A. Discussion of Findings

Research on echo chambers in recommender systems has consistently highlighted their pervasive influence across various recommendation algorithms, including collaborative, content-based, and hybrid systems. A major limitation of collaborative filtering, for example, is its inherent reliance on past user behavior and preferences of similar users. This often leads to feedback loops where users are repeatedly exposed to the same type of content, reinforcing their preferences and restricting their exposure to diverse perspectives. This absence of diversity keeps users in echo chambers, narrowing their content exploration and contributing to polarization [98].

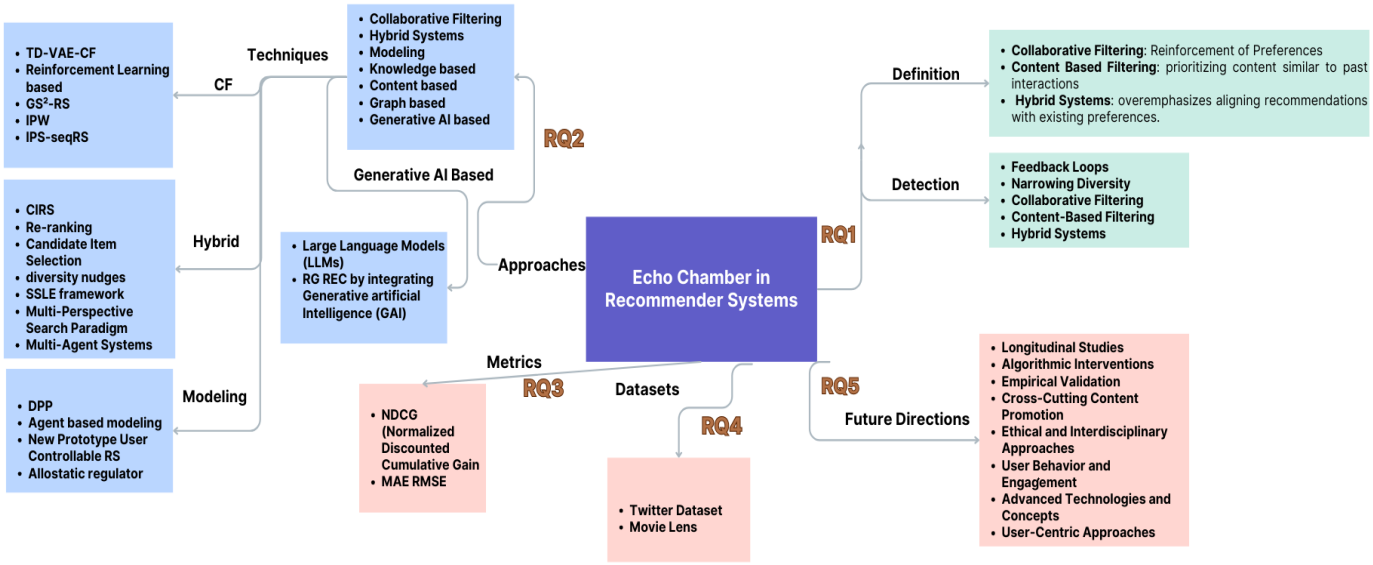


Fig. 7. Overview of findings for the research questions.

To address these limitations, several innovative approaches have been proposed. Targeted Diversification VAE-based Collaborative Filtering (TD-VAE-CF) [60] and Reinforcement Learning-based methods mitigate the feedback loop issue by introducing varied recommendations, diversifying the range of topics and content suggested to users. However, these methods often face challenges in maintaining relevance while enhancing diversity, as balancing the two can sometimes lead to reduce user satisfaction. Community-Aware Models and Contrarian Recommendations attempt to address these issues by integrating social context and intentionally suggesting items that contrast with user preferences. These models break the echo chamber cycle by exposing users to diverse viewpoints and encouraging engagement with content [55]. More advanced approaches, such as the Triple-Filter-Bubble framework utilizing an Agent-Based Model (ABM) [103] tackle systemic challenges like societal fragmentation. ABMs simulate cognitive processes behind echo chambers, including confirmation bias, offering valuable insights into how personalized recommendations contribute to polarization. Yet, their reliance on complex simulations may limit their applicability in real-time systems. Agent4Rec "a recommendation simulator", extends this work by addressing unresolved challenges like causality and user engagement. This approach highlights the importance of understanding the interplay between user behavior and recommendation algorithms, but its use in real-world applications remains questionable [63].

Another valuable direction focusing on maximizing semantic volume and linguistic diversity to enrich content recommendations. Expanding the range of topics and styles, these systems reduce bias and enhance user exposure to new perspectives. RGRec, combining graph-based reasoning, belief nudging, and generative AI, provides an advanced solution by detecting imbalances in user beliefs and nudging users toward diverse content [66]. While the system effectively encourages greater exploration, its reliance on belief manipulation raises ethical concerns about user autonomy. Various scales were used to evaluate these techniques. Diversity Metrics (e.g., ILD, Gini-Simpson Index) are critical for assessing how well systems expose users to a wide variety of content. However, over-reliance on traditional metrics like Precision and Recall, which prioritize accuracy over diversity, can inadvertently reinforce echo chambers. Balancing these metrics with diversity measures remains a key challenge.

Datasets such as MovieLens and Twitter are frequently used to study echo chambers, offering insights into their development across different platforms. However, these datasets primarily focus on entertainment and social media domains, limiting generalizability to contexts like e-commerce or news. Expanding research to incorporate diverse datasets is essential for understanding echo chambers across various sectors.

In summary, while existing methods face challenges such as balancing relevance and diversity or ensuring real-time applicability, newer approaches like graph-based reasoning, semantic diversity, and agent-based simulations offer promising solutions. By addressing these limitations and applying a combination of innovative techniques with evaluation metrics, researchers can more effectively mitigate echo chambers. The continued use of diverse datasets will help ensure that these strategies remain robust and adaptable across multiple domains [40], [66], [83].

B. Limitations of this Systematic Literature Review

A limitation of this review is that, although it covers studies from 2019 to 2025, the analysis only incorporates work published up to early 2025. Any significant research released after this point has not been captured, which may limit the completeness of the findings. Furthermore, despite selecting studies from six major digital libraries, there may still be relevant research from other sources that was not included. The review also focused exclusively on English-language publications, meaning that important non-English studies might

have been overlooked. Additionally, due to time constraints, this review relies on self-reported performance results from the authors of the selected studies, which may introduce bias into the assessment.

VII. Conclusion and Future Works

In this SLR, we have selected 56 research papers, published between 2019 and 2025, to explore how echo chambers appear in recommender systems and the solutions proposed to mitigate them. The main strength of this review is the use of a clear and systematic methodology, including a wide search strategy, predefined keywords, and strict inclusion and exclusion criteria. This ensures that the findings are both reliable and comprehensive. The study not only highlights the influence of echo chambers on user experience and content diversity but also evaluates the effectiveness of existing approaches to reduce polarization. Therefore, this work is valuable for guiding both researchers and practitioners, and it can also support the development of recommender systems that balance personalization with exposure to diverse content. Despite these contributions, this review has several limitations. The analysis focused primarily on studies published in English and indexed in major databases, which may exclude relevant non-English or gray literature. In addition, many included studies relied on simulated or single-platform datasets, limiting the generalizability of conclusions about real-world systems.

As a future work, we plan to build this SLR to address these limitations by conducting cross-platform, longitudinal, and user-centered analyses to better capture the dynamic nature of echo chambers. Promising directions include the integration of natural language processing, state space models, and other advanced AI techniques to design adaptive recommenders that maintain personalization while promoting exposure to diverse and balanced content.

References

[1] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An open architecture for collaborative filtering of Netnews," in Proc. ACM Conf. Computer Supported Cooperative Work (CSCW), Chapel Hill, NC, USA, 1994, pp. 175–186.

[2] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," Computer, vol. 42, no. 8, pp. 30–37, Aug. 2009, doi: 10.1109/MC.2009.263.

[3] A. Töscher, M. Jahrer, and R. M. Bell, "The BigChaos solution to the Netflix Grand Prize," *Netflix Prize Documentation*, 2009, pp. 1–52.

[4] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for YouTube recommendations," in *Proc. ACM Conf. Recommender Systems (RecSys)*, 2016.

[5] F. Sun, L. Yao, and Y. Li, "BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformers," in *Proc. ACM Conf. Recommender Systems (RecSys)*, 2019.

[6] J. Zhao, H. Li, L. Qu, Q. Zhang, Q. Sun, H. Huo, and M. Gong, "DCFGAN: An adversarial deep reinforcement learning framework with improved negative sampling for session-based recommender systems," *Information Sciences*, vol. 596, pp. 222–235, 2022.

[7] P. Törnberg, D. Valeeva, J. Uitermark, and C. Bail, "Simulating social media using large language models to evaluate alternative news feed algorithms," *arXiv preprint*, arXiv:2310.05984, 2023.

[8] C. C. Aggarwal, "A comprehensive survey on recommender systems," 2018.

[9] M. Cinelli, A. Galeazzi, S. Galea, and C. Valensise, "The echo chamber effect on social media: A study of the role of recommender algorithms," *Journal of Computational Social Science*, 2020.

[10] P. Barberá, "How social media reduces mass political polarization: Evidence from Germany, Spain, and the U.S.," 2014.

[11] J. Kim and M. Kim, "The effect of personalized recommendations on consumer behavior: An empirical study of e-commerce platforms," *Electronic Commerce Research and Applications*, 2017.

[12] C. T. Nguyen, "Cognitive islands and runaway echo chambers: Problems for expert dependence," *Synthese*, 2018. doi: 10.1007/s11229-018-1692-0.

[13] C. Vaccari and A. Valeriani, "Echo chambers in online news consumption: Evidence from an experimental study," *Journal of Information Technology & Politics*, 2021.

[14] C. R. Sunstein, *Republic.com 2.0*, Princeton, NJ, USA: Princeton University Press, 2007.

[15] E. Pariser, *The Filter Bubble: What the Internet is Hiding from You*, New York, NY, USA: Penguin Press, 2011.

[16] J. Möller, D. Trilling, N. Helberger, and B. van Es, "Do not blame it on the algorithm: An empirical assessment of multiple recommender systems and their impact on content diversity," *Information, Communication & Society*, vol. 21, no. 7, pp. 959–977, 2018.

[17] C. C. Aggarwal, "A comprehensive survey of content-based recommender systems: Techniques and applications," *ACM Computing Surveys (CSUR)*, 2016.

[18] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proc. 10th Int. Conf. World Wide Web (WWW)*, 2001, doi: 10.1145/371920.372071.

[19] Y. Koren and R. Bell, "Matrix factorization techniques for recommender systems," *Computer*, vol. 38, no. 8, pp. 30–37, 2003, doi: 10.1109/MC.2005.263.

[20] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques with social regularization for collaborative filtering," in Proc. ACM RecSys, 2010, doi: 10.1145/1864708.1864727.

[21] R. Burke, "A survey of hybrid recommender systems: Techniques and applications," *ACM Computing Surveys (CSUR)*, 2002, doi: 10.1145/582034.582036.

[22] X. Liang, Z. Li, and F. A. Wei, "Hybrid recommender system based on sentiment analysis and matrix factorization," in *Proc. Int. Joint Conf. Artificial Intelligence (IJCAI)*, 2020.

[23] P. Lops, M. de Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender Systems Handbook*, Springer, 2011, pp. xx–xx.

[24] K. Garrett, "Echo chambers online? Politically motivated selective exposure among Internet news users," *J. Comput.-Mediat. Commun.*, vol. 14, no. 2, pp. 265–285, Jan. 2009.

[25] C. R. Sunstein, "The law of group polarization," *J. Political Philosophy*, vol. 10, no. 2, pp. 175–195, Jun. 2002.

[26] R. Jiang, S. Chiappa, T. Lattimore, A. György, and P. Kohli, "Degenerate feedback loops in recommender systems," in *Proc. AAAI/ACM Conf. AI, Ethics, and Society (AIES)*, New York, NY, USA, 2019, pp. 383–390, doi: 10.1145/3306618.3314288.

[27] W. Wang, F. Feng, L. Nie, and T.-S. Chua, "User-controllable recommendation against filter bubbles," in *Proc. 45th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 2022.

[28] M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, "The echo chamber effect on social media," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 118, no. 9, Mar. 2021.

[29] J. Sun, J. Song, Y. Jiang, Y. Liu, and J. Li, "Prick the filter bubble: A novel cross-domain recommendation model with adaptive diversity regularization," *Electronic Markets*, pp. 1–21, 2021.

[30] B. Kitchenham and P. Brereton, "A systematic review of systematic review process research in software engineering," *Information and Software Technology*, vol. 55, no. 12, pp. 2049–2075, 2013.

[31] R. Interian, G. Marzo, R. Mendoza, and C. C. Ribeiro, "Network polarization, filter bubbles, and echo chambers: An annotated review of measures and reduction methods," Int. Trans. Oper. Res., vol. 30, no. 6, pp. 3122–3158, 2022.

[32] X. Ma, Q. Hu, Z. Gao, and M. AbdelHady, "Contrastive co-training for diversified recommendation," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), 2022.

[33] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis, "Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship," in Proc. 2018 World Wide Web Conf. (WWW '18), Geneva, Switzerland: IW3C2, 2018, pp. 913–922.

[34] M. J. Alam, I. Hossain, S. Puppala, and S. Talukder, "Combating echo chambers in online social network by increasing content diversity in recommendation," in Proc. Int. Conf., Sep. 2024, pp. 1–16

[35] C. Zhang, Z. Li, and K. Zhang, "An inverse propensity score framework for breaking filter bubble recommendation," in Proc. IEEE Int. Conf. Control, Electron. Comput. Technol. (ICCECT), 2023.

[36] F. Cinus, M. Minici, C. Monti, and F. Bonch, "The effect of people recommenders on echo chambers and polarization," 2022.

[37] Y. Gao, Z. Wang, J. Li, and L. Zhang, "Echo Chamber Effects on Short Video Platforms," Proc. Int. Conf. on Web and Social Media (ICWSM), 2023.

[38] Z. Li et al., "Breaking filter bubble: A reinforcement learning framework of controllable recommender system," in Proc. ACM Web Conf., 2023.

[39] C. Gao et al., "CIRS: Bursting filter bubbles by counterfactual interactive recommender system," ACM Trans., 2023.

[40] A. Vendeville, A. Giovanidis, E. Papanastasiou, and B. Guedj, "Opening up echo chambers via optimal content recommendation," in Complex Networks and Their Applications XI, H. Cherifi, R. N. Mantegna, L. M. Rocha, C. Cherifi, and S. Miccichè, Eds. Cham: Springer, 2023, vol. 1077, Studies in Computational Intelligence, pp. 83–94. doi: 10.1007/978-3-031-21127-0_7.

[41] I. B. Dylko, M. A. Beam, K. D. Landreville, and N. Geidner, "Echo chambers in the age of algorithms: Evaluating the impact of personalized news recommendations on ideological isolation," J. Inf. Technol. Politics, vol. 15, no. 2, pp. 81–94, 2018, doi: 10.1080/19331681.2018.1448735.

[42] W. Cota, S. C. Ferreira, R. Pastor-Satorras, and M. Starnini, "Quantifying echo chamber effects in information spreading over political communication networks," EPJ Data Sci., vol. 8, no. 1, pp. 1–13, Dec. 2019.

[43] V. R. K. Garimella and I. Weber, "A long-term analysis of polarization on Twitter," in Proc. Int. Conf. Weblogs and Social Media (ICWSM), vol. 11, no. 1, pp. 528–531, May 2017.

[44] J. K. Lee, J. Choi, C. Kim, and Y. Kim, "Social media, network heterogeneity, and opinion polarization," J. Commun., vol. 64, no. 4, pp. 702–722, 2014.

[45] D. Nikolov, A. Flammini, and F. Menczer, "Right and left, partisanship predicts (asymmetric) vulnerability to misinformation," Harvard Kennedy School Misinformation Rev., 2021. [Online]. Available: https://doi.org/10.37016/mr2020-55

[46] L. Terren and R. Borge-Bravo, "Echo chambers on social media: A systematic review of the literature," Rev. Com. Rhetor., vol. 9, pp. 99–118, Mar. 2021.

[47] E. Dubois and G. Blank, "The echo chamber is overstated: The moderating effect of political interest and diverse media," Inf. Commun. Soc., vol. 21, no. 5, pp. 729–745, May 2018.

[48] E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on Facebook," Science, vol. 348, no. 6239, pp. 1130–1132, 2015.

[49] A. Gharahighehi and C. Vens, "Diversification in session-based news recommender systems," Pers. Ubiquitous Comput., pp. 1–11, 2021.

[50] M. Zhitomirsky-Geffet, "Turning filter bubbles into bubblesphere with multi-viewpoint KOS and diverse similarity," in Proc. Assoc. Inf. Sci. Technol., 2022.

[51] G. M. Lunardi, "Representing the filter bubble: Towards a model to diversification in news," Lect. Notes Comput. Sci., vol. 2019, 2019.

[52] A. Mahmoudi, D. Jemielniak, and L. Ciechanowski, "Echo chambers in online social networks: A systematic literature review," IEEE Access, vol. 12, pp. 9594–9620, 2024, doi: 10.1109/ACCESS.2024.3353054.

[53] Q. M. Areeb et al., "Filter bubbles in recommender systems: fact or fallacy—a systematic review," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 13, no. 6, e1512, 2023, doi: 10.1002/WIDM.1512.

[54] N. Henry, M. Pedersen, M. Williams, J. Martin, and L. Donkin, "Reducing echo chamber effects: An allostatic regulator for recommendation algorithms," 2023.

[55] D. Geschke, J. Lorenz, and P. Holtz, "The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers," Br. J. Soc. Psychol., 2019.

[56] B. Paudel and A. Bernstein, "Random walks with erasure: Diversifying personalized recommendations on social and information networks," in Proc. Web Conf., 2021.

[57] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," Annu. Rev. Sociol., vol. 27, no. 1, pp. 415–444, 2001.

[58] Y. Ge, S. Zhao, H. Zhou, C. Pei, F. Sun, W. Ou, and Y. Zhang, "Understanding echo chambers in e-commerce recommender systems," in Proc. 43rd Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, 2020, pp. 2261–2270.

[59] H. Zhang, Z. Zhu, and J. Caverlee, "Evolution of filter bubbles and polarization in news recommendation," in Proc. Eur. Conf. Inf. Retrieval, 2023.

[60] Z. Gao, T. Shen, Z. Mai, M. Bouadjenek, I. Waller, A. Anderson, R. Bodkin, and S. Sanner, "Mitigating the filter bubble while maintaining relevance: Targeted diversification with VAE-based recommender systems," in Proc. 45th Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, 2022, doi: 10.1145/3477495.3531890.

[61] N. Sukiennik, C. Gao, and N. Li, "Uncovering the Deep Filter Bubble: Narrow Exposure in Short-Video Recommendation," in Proc. ACM Web Conf. (WWW '24), 2024, pp. 4727–4735, doi: 10.1145/3589334.3648159.

[62] K. Duskin, J. S. Schafer, J. D. West, and E. S. Spiro, "Echo Chambers in the Age of Algorithms: An Audit of Twitter's Friend Recommender System," in Proc. 16th ACM Web Science Conf. (WebSci '24), 2024, pp. 11–21, doi: 10.1145/3614419.3643996.

[63] A. Zhang, Y. Chen, L. Sheng, X. Wang, and T. Chua, "On Generative Agents in Recommendation," in Proc. 47th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR '24), 2024, pp. 1807–1817, doi: 10.1145/3626772.3657844.

[64] A. Starke, A. S. Bremnes, E. Knudsen, D. Trilling, and C. Trattner, "Perception versus Reality: Evaluating User Awareness of Political Selective Exposure in News Recommender Systems," in Adjunct Proc. 32nd ACM Conf. on User Modeling, Adaptation and Personalization (UMAP Adjunct '24), 2024, pp. 286–291, doi: 10.1145/3631700.3665189.

[65] Y. Xu, E. Wang, Y. Yang, and H. Xiong, "GS-RS: A Generative Approach for Alleviating Cold Start and Filter Bubbles in Recommender Systems," IEEE Trans. Knowl. Data Eng., vol. 36, no. 2, pp. 668–681, Feb. 2024, doi: 10.1109/TKDE.2023.3290140.

[66] M. Wang et al., "Nudging Towards Responsible Recommendations: A Graph-Based Approach to Mitigate Belief Filter Bubbles," IEEE Trans. Artif. Intell., 2024, doi: 10.1109/TAI.2024.3373392.

[67] J. Duskin and B. Schäfer, "Echo Chambers in the Age of Algorithms: An Audit of Twitter's Who-to-Follow," Proc. ACM Conf. on Web Science (WebSci), 2024.

[68] Md. Anwar, G. Schoenebeck, and P. S. Dhillon, "Filter Bubble or Homogenization? Disentangling the Long-Term Effects of Recommendations on User Consumption Patterns," in Proc. ACM Web Conf. (WWW '24), 2024, pp. 123–134, doi: 10.1145/3589334.3645497.

[69] D. Trilling, "Communicative Feedback Loops in the Digital Society," Weizenbaum Journal of the Digital Society, vol. 4, no. 2, 2024, doi: 10.34669/wi.wjds/4.2.4.

[70] N. Mattis, P. Masur, J. Möller, and W. van Atteveldt, "Nudging Towards News Diversity: A Theoretical Framework for Facilitating Diverse News Consumption through Recommender Design," New Media & Society, vol. 26, no. 7, pp. 3681–3706, 2024.

[71] N. Sharma, Q. V. Liao, and Z. Xiao, "Generative echo chamber? Effect of LLM-powered search systems on diverse information seeking," in Proc. CHI Conf. on Human Factors in Computing Systems (CHI), 2024, pp. 1–17.

[72] Y. Zheng, Z. Chen, J. Qin, and L. Lin, "FacetCRS: Multi-Faceted Preference Learning for Pricking Filter Bubbles in Conversational Recommender System," in Proc. AAAI Conf. Artif. Intell., vol. 38, no. 8, 2024, doi: 10.1609/aaai.v38i8.28794.

[73] M. Iwanaga, K. Sato, and T. Nakajima, "Echo-Chamber Formation by Friend Recommendation Algorithms," Proc. ACM Conf. on Recommender Systems (RecSys), 2025.

[74] X. Zhou et al., "Bee and I Need Diversity! Break Filter Bubbles in Recommendation Systems through Embodied AI Learning," in Proc. 23rd ACM Interaction Design and Children Conf. (IDC '24), 2024, pp. 44–61, doi: 10.1145/3628516.3655802.

[75] S. Nagraj and B. Palayyan, "Personalized E-commerce Based Recommendation Systems Using Deep-Learning Techniques," Int. J. Artif. Intell., vol. 13, no. 1, pp. 610–618, Mar. 2024, doi: 10.11591/ijai.v13.i1.pp610-618.

[76] E. Rodilosso, "Filter Bubbles and the Unfeeling: How AI for Social Media Can Foster Extremism and Polarization," Philos. Technol., vol. 37, art. 71, 2024, doi: 10.1007/s13347-024-00758-4.

[77] J. Y. Sun et al., "See Widely, Think Wisely: Toward Designing a Generative Multi-Agent System to Burst Filter Bubbles," in Proc. CHI Conf. Human Factors in Computing Systems (CHI '24), 2024, Art. no. 484, pp. 1–24, doi: 10.1145/3613904.3642545.

[78] H. Han, C. Wang, Y. Zhao et al., "SSLE: A Framework for Evaluating the 'Filter Bubble' Effect on the News Aggregator and Recommenders," World Wide Web, vol. 25, pp. 1169–1195, 2022, doi: 10.1007/s11280-022-01031-4.

[79] T. Donkers and J. Ziegler, "De-sounding Echo Chambers: Simulation-Based Analysis of Polarization Dynamics in Social Networks," 2023. [Online]. Available: [Exact source or DOI missing].

[80] S. Vrijenhoek and N. Helberger, "Recommenders with a Mission – Assessing Diversity in News Recommendations," in Proc. ECML PKDD 2020 Workshops, 2020.

[81] Q. Grossetti, C. du Mouza, and N. Travers, "Community-Based Recommendations on Twitter: Avoiding the Filter Bubble," Lect. Notes Comput. Sci., vol. 11881, pp. 212–227, 2019.

[82] M. El Khoury Badran, J. B. Abdo, W. Jurdi, and J. Demerjian, "Adaptive Serendipity for Recommender Systems: Let It Find You," in Proc. 11th Int. Conf. Agents and Artificial Intelligence (ICAART), vol. 2, pp. 739–745, 2019.

[83] Sonoda, F. Toriumi, and H. Nakajima, "Improving Recommendation Diversity by Maximizing Semantic Volume," in Proc. 2022 IEEE/WIC/ACM Int. Joint Conf. Web Intelligence and Intelligent Agent Technology (WI-IAT), Niagara Falls, ON, Canada, 2022, pp. 496–503, doi: 10.1109/WI-IAT55865.2022.00079.

[84] M. Alam et al., "Towards Analyzing the Bias of News Recommender Systems Using Sentiment and Stance Detection," in Companion Proc. Web Conf. (WWW), 2022, pp. 448–457.

[85] A. Vivek, M. Yang, and Z. Zhao, "Mitigating Filter Bubbles within Deep Recommender Systems," in Data Science for Social Networks, Atlanta, GA, Dec. 2021. [Online]. Available: http://arxiv.org/abs/2209.08180

[86] F. Bagnoli, G. de Bonfioli Cavalcabo, B. Casu, and A. Guazzini, "Bubble Effect Induced by Recommendation Systems in a Simple Social Media Model," Stud. Comput. Intell., vol. 1016, pp. 124–131, 2022.

[87] Y. Ge, S. Zhao, H. Zhou, C. Pei, F. Sun, W. Ou, and Y. Zhang, "Deconstructing the Filter Bubble: User Decision-Making and Recommender Systems," in Companion Proc. Web Conf. (WWW), 2020, pp. 448–457.

[88] S. Khenissi, M. Boujelbene, and O. Nasraoui, "Theoretical Modeling of the Iterative Properties of User Discovery in a Collaborative Filtering Recommender System," in Proc. 14th ACM Conf. Recommender Systems (RecSys '20), 2020.

[89] D. Kalimeris, S. Bhagat, S. Kalyanaraman, and U. Weinsberg, "Preference Amplification in Recommender Systems," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 805–815, 2021.

[90] Q. Grossetti, C. Du Mouza, N. Travers, and C. Constantin, "Reducing the Filter Bubble Effect on Twitter by Considering Communities for Recommendations," Int. J. Web Inf. Syst., vol. 17, no. 6, pp. 728–752, 2021

[91] A. De Biasio, M. Monaro, L. Oneto, L. Ballan, and N. Navarin, "On the problem of recommendation for sensitive users and influential items: Simultaneously maintaining interest and diversity," Knowl.-Based Syst., 2023.

[92] H. Xia, K. Huang, and Y. Liu, "Unexpected interest recommender system with graph neural network," Complex Intell. Syst., 2023.

[93] L. Michiels, J. Leysen, A. Smets, and B. Goethals, "What are filter bubbles really? A review of the conceptual and empirical work," in Adjunct Proc. 30th ACM Conf. User Modeling, Adaptation and Personalization (UMAP '22 Adjunct), New York, NY, USA: ACM, 2022, pp. 274–279.

[94] Y. Peng, X. Qian, and W. Song, "A Re-ranking Approach for Two-sided Fairness on Recommendation Systems," ACM Int. Conf. Proc. Ser., 2023.

[95] JJ. Wu, B. Yang, R. Mao, and Q. Li, "Popularity-aware sequential recommendation with user desire," Expert Syst. Appl., vol. 237, 2024, doi: 10.1016/j.eswa.2023.121429..

[96] F. V. Vromman and F. Fouss, "Filter bubbles created by collaborative filtering algorithms themselves, fact or fiction? An experimental comparison," ACM Int. Conf. Proc. Ser., 2021.

[97] G. M. Lunardi, G. M. Machado, V. Maran, and J. P. M. de Oliveira, "A metric for Filter Bubble measurement in recommender algorithms considering the news domain," Appl. Soft Comput. J., 2020.

[98] A. Godinot and F. Tarissan, "Measuring the effect of collaborative filtering on the diversity of users' attention," Appl. Netw. Sci., vol. 8, no. 1, p. 9, 2023.

[99] A. Tommasel, J. M. Rodriguez, and D. Godoy, "I Want to Break Free! Recommending Friends from Outside the Echo Chamber," in Proc. 15th ACM Conf. Recommender Syst. (RecSys), 2021, pp. 23–33.

[100] R. Qin, B. Mac Namee, and R. Dong, "The Influence of Media Bias on News Recommender Systems," in Proc. 31st ACM Conf. User Modeling, Adaptation and Personalization (UMAP '23), 2023, pp. 301–305, doi: 10.1145/3565472.3595619.

[101] L. Pappalardo et al., "A survey on the impact of AI-based recommenders on human behaviours: methodologies, outcomes and future directions," arXiv preprint arXiv:2407.01630, 2024.

[102] S. C. Matz, "Personal echo chambers: Openness-to-experience is linked to higher levels of psychological interest diversity in large-scale behavioral data," J. Pers. Soc. Psychol., vol. 121, no. 6, pp. 1284–1300, 2021, doi: 10.1037/pspp0000324.

[103] T. Donkers and J. Ziegler, "The Dual Echo Chamber: Modeling Social Media Polarization for Interventional Recommending," 2021.

[104] Y. Jeon, B. Kim, A. Xiong, D. Lee, and K. Han, "Chamber Breaker: Mitigating the Echo Chamber Effect and Supporting Information Hygiene through a Gamified Inoculation System," 2021.

[105] V. S. Chaitanya, M. Mohan and P. S. Thilagam, "A Clustering-based model for the Generation of Diversified Recommendations," 2022 IEEE Silchar Subsection Conference (SILCON), Silchar, India, 2022, pp. 1–6, doi: 10.1109/SILCON55242.2022.10028791.

[106] L. Bojic, "A Novel Model for Diversifying AI-Based Recommender Systems for Societal Well-Being," ESWC 2025 Workshops and Tutorials Joint Proceedings, CEUR Workshop Proceedings, 2025, pp. 1–15

[107] M. C. Cakmak and N. Agarwal, "Influence of symbolic content on recommendation bias: analyzing YouTube's algorithm during Taiwan's 2024 election," Applied Network Science, vol. 10, no. 23, 2025, doi: 10.1007/s41109-025-00713-y.

[108] F. Lomonaco, D. Taibi, V. Trianni, G. Donabauer, and D. Ognibene, "Yes, echo-chambers mislead you too: A game-based educational experience to reveal the impact of social media personalization algorithms," in Commun. Comput. Inf. Sci. (CCIS), vol. 1779, pp. 330–344, 2023.

Critical Review of Object Detection Techniques for Traffic Light Detection in Intelligent Transportation Systems

Adhwa Salemi, Muhammad Arif Mohamad
Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan, Malaysia

***Abstract*—Object detection and tracking play a critical role in intelligent transportation systems (ITS), particularly in recognizing and monitoring traffic lights to ensure safety and improve traffic efficiency. Despite progress in deep learning and optimization algorithms, traffic light detection still faces persistent challenges under varying conditions such as illumination changes, occlusions, and visual clutter. This study provides a critical review of object detection techniques specifically for traffic light detection, evaluating the evolution of machine learning frameworks, deep learning architectures, and hybrid optimization models. The review identifies research gaps in the robustness, real-time adaptability, and generalizability of existing methods. Furthermore, it highlights emerging trends such as multi-camera systems, anchor-free detection, and hybrid optimization techniques that bridge performance trade-offs between accuracy and efficiency. The findings offer a new perspective on integrating multiple approaches to achieve scalable, high-accuracy traffic light detection for future ITS applications.**

***Keywords*—Object detection; traffic light detection; optimization; intelligent transportation systems; review**

I. Introduction

The increasing use of intelligent transportation systems (ITS) and autonomous vehicles has dramatically altered traffic management and road safety paradigms [1]. Several enabling elements of such innovations, such as traffic light detection and tracking, are of pivotal importance. Traffic light signals form the backbone of orderly traffic flow and avoidance of collisions, and hence their detection and tracking with high accuracy is of critical importance to efficient ITS operations and autonomous vehicle capabilities [2].

In spite of their significance, the utilization of traffic light detection systems in real-world settings is beset with difficulties [3]. Such difficulties encompass complicated city landscapes, occlusions caused by other vehicles and roadside installations, variations in the color and pattern of traffic lights across different regions, in addition to the impact of adverse weather conditions such as low lighting, rain, and fog [4]. For example, the significance of deep models for addressing environmental challenges is best illustrated through their ability to detect traffic lights in challenging lighting conditions [5]. These challenges not only compromise detection quality but also hinder smooth tracking between frames, a fundamental element of autonomous decision-making.

Although several review papers have discussed object detection in general contexts, few have specifically addressed the unique challenges of traffic light detection, especially under environmental variations like low illumination, occlusion, and cluttering. Moreover, most existing reviews focus either on traditional image processing or deep learning independently, without synthesizing insights across optimization algorithms and real-time deployment frameworks. This gap highlights the need for a comprehensive review that bridges these areas to support more robust, real-time ITS solutions.

To address these issues, researchers have explored a range of object-detection architectures and optimization methods designed for complex real-world traffic environments. Object detection is a fundamental task within computer vision that enables systems to identify and locate objects within images or video feeds [6]. Traditional approaches relied on hand-designed features, including Haar cascades and histograms of oriented gradients, but the accuracy of detection has significantly increased since the introduction of deep learning techniques [7]. Convolutional neural networks (CNNs) form the foundation of object detection in modern computer vision, wherein the Faster R-CNN and YOLO models are particularly well adapted for real-time use [8]. Vision transformers are being incorporated into detection pipelines through self-attention mechanisms to boost feature extraction [9]. Multi-modal sensor fusion methods that integrate visual with LiDAR or radar data have also increased detection robustness in autonomous systems [10]. Precision, recall, and intersection over union (IoU) metrics remain significant in evaluating model performance [11].

Recent research has also studied progressive multi-modal fusion methods in which object detection is aided through hierarchical feature combination in various sensor modalities [12]. Researchers have also studied the effects of mono-modality feature learning in multi-modal object detection in resolving problems of degradation in fusion [13]. Development of optimal performance assessment frameworks has further improved standards of benchmarking object detection models [14].

As research develops, object detection becomes more advanced with the use of improved architectures and optimisation methods to achieve improved accuracy and efficiency [15]. Growing demand for reliable solutions has driven ample research in traffic light detection and tracking

methods. Conventional methods that include rule-based methods and traditional computer vision methods have been extensively employed; nonetheless, these cannot adequately cope with dynamic and complex environments [16].

Accordingly, this study aims to provide a targeted and critical survey of traffic light detection techniques. The discussion is structured as follows: Section II reviews past and current detection methods, including hybrid and optimization-based approaches. Section III explores recent trends in deep learning and real-time processing frameworks. Section IV evaluates these trends through comparative metrics and radar analysis, and Section V concludes with proposed research directions for more adaptive, scalable detection systems.

II. CRITICAL SURVEY

The studies reviewed in this section primarily utilize diverse datasets such as COCO, BDD100K, and custom traffic-light datasets collected under varied environmental conditions (day/night/rain/fog), ensuring that comparisons account for both real-world and synthetic data sources.

As city life is getting more complex, accurate and efficient traffic-light detection is crucial for maintaining traffic flow and ensuring road safety. Researchers initially used rule-based algorithms and traditional computer vision methods with a focus on colour segmentation, edge detection, and template matching to detect traffic lights from recorded images and video clips [17]. These methods worked well in ideal conditions, but failed in changing conditions. False detection and low confidence typically resulted from challenges such as changing light conditions, various traffic light configurations, and occlusions [18].

The deep learning era has changed how we detect objects, like traffic lights, with more effective models compared to conventional methods. Convolutional neural networks (CNNs) can detect complex features and are the foundation of modern detection systems. Among them, You Only Look Once (YOLO) models like YOLOv4, YOLOv5, and YOLOv7 have excelled in real-time applications. Experiments show how YOLO works together with tracking methods like Kalman filters and DeepSORT in achieving high accuracy for traffic monitoring systems [19]. Similarly, YOLO-based model research highlights their possibilities in solving problems caused by low light and short-time blockages [20] [15].

Mono and hybrid methods that combine optimization algorithms and deep learning have been used intensively in the last several years. These approaches take advantage of the strengths of the two paradigms to improve object detection systems. For example, the Firefly Algorithm (FA) is based on the behaviour of the fireflies [21], to enhance detection performance in difficult conditions. One notable example is KCS-YOLO, an improved YOLOv5n variant that utilizes attention and clustering mechanisms to enhance detection under low-visibility scenes [22]. The Zebra Optimization Algorithm (ZOA) is also a promising optimization algorithm that has been shown to achieve identical and continuous recognition in different video frames with its effectiveness in complex traffic scenes [23].

The demand for real-time processing in ITS applications has also promoted innovations in deployment frameworks. Frameworks such as TensorRT and ONNX streamline the deployment of deep learning models in edge devices to achieve efficient traffic light detection and tracking in real-time applications. For instance, TensorRT is used to speed up YOLOv5-based systems to balance speed and accuracy in autonomous vehicles [15]. Such frameworks are important in helping to meet the computational requirements of ITS applications to allow scalable and robust detection systems.

Recent studies demonstrate the promising work being done in this area. Table I below indicates the total papers that aim at object detection, particularly in traffic signals.

TABLE I. APPROACHES FOR TRAFFIC LIGHT DETECTION IN RECENT YEARS

AUTHOR (YEAR)	TYPE OF APPROACHES
De Guia et al. (2023) [24]	Utilized YOLOv7 for real-time traffic light detection, focusing on improving accuracy under occlusions and poor lighting.
Wu et al. (2023) [15]	Proposed a multi-camera system with YOLOv5 for robust detection and tracking, particularly addressing flashing light states.
Zhou et al. (2023) [22]	Introduced KCS-YOLO, an optimized YOLOv5n model with attention mechanisms to enhance performance under low visibility.
Vedant Singh (2025) [19]	Combined YOLO and SSD with tracking algorithms like DeepSORT for real-time traffic surveillance systems.
Bochkovskiy et al. (2020) [25]	Integrated Firefly Algorithm with YOLOv4 for optimizing object detection in low-light conditions.
Liu et al. (2024) [26]	Developed a GAN-based synthetic dataset for training object detection models, improving generalization for varied traffic environments.
Talaat et al. (2023) [27]	Adopted SSD with Kalman filters for tracking, achieving robust detection across consecutive frames.
Nagendra Rao et al. (2024) [28]	Explored transfer learning using ResNet50, focusing on adapting object tracking models for diverse regional traffic light designs
Chen et al. (2021) [29]	Applied single-shot detector (SSD) models optimized with IoU-based loss functions for better tracking of small objects like traffic lights
Zhang et al. (2021) [30]	Proposed an anchor-free detection technique to simplify tracking and improve inference speed in dense urban traffic scenarios

In spite of the advancements in object detection towards traffic light detection based on the papers above, there are various challenges that persist. Occlusions, illumination, and cluttering continue to plague the detection systems and call for more reliable solutions [24]. Moreover, diversity in traffic light configurations and designs in various regions challenges model flexibility and generalizability [25]. Future work will be required to utilize models that combine multiple algorithms or deep learning for the best solution and investigate the use of synthetic datasets and transfer learning to broaden system flexibility [26][27].

Unlike previous reviews that focus solely on specific algorithms or environments, this study synthesizes cross-disciplinary insights combining deep learning, optimization algorithms, and hardware deployment perspectives. This integrative view positions the study not merely as a compilation, but as an analytical framework for identifying

hybrid approaches that balance accuracy, adaptability, and computational cost.

III. Current Trends in Object Detection for Traffic Light Detection

The research in object detection towards traffic light detection has been significantly improved in recent years with the growing demands of intelligent traffic systems and autonomous vehicles [16]. Current trends manifest in a move to utilize state-of-the-art technologies to overcome challenges in dynamic traffic scenes, variable lighting conditions, and real-time processing requirements [31]. One of the most prominent trends is the use of anchor-free detection models. In contrast to conventional anchor-based methods that use prior bounding box sizes, anchor-free models streamline detection in that they predict object locations and sizes directly [32]. This reduces computational complexity and ensures speed in detection and is hence most applicable in real-time applications. One such example is an anchor-free detection method called DARDet that was proposed to enhance inference speed and accuracy in dense urban traffic environments [30].

Another recent development involves incorporating attention mechanisms into object detection models. They enable the models to focus on the most prominent features in an image, improving detection accuracy even in challenging situations like occlusion or poor lighting. An example is KCS-YOLO, a variant of YOLOv5n that incorporates attention mechanisms to improve detection accuracy in low-light conditions, with significant improvement in tracking precision [22].

Integration of multi-camera systems is gaining traction in traffic light detection. Multi-camera configurations possess a larger field of view and enable the merging of multi-view information, thereby rendering it more precise and robust. A proposed instance is a multi-camera traffic light detection system based on the synergistic application of YOLOv5 along with merged high-definition semantic maps and hidden Markov models. This aims to allow for complicated scenarios, like numerous intersections and various flashing light modes, and more precise outcomes in dynamic environments [15].

Hybrid approaches that combine deep learning and metaheuristic optimization approaches are one of the biggest future trends [35][36]. Such approaches take advantage of the best in the two paradigms to eliminate the limitations of traditional methods. For example, the Firefly Algorithm, inspired by the behaviour of fireflies, has been combined with YOLO-based methods to increase detection accuracy and tracking stability in adverse conditions. Equally, one study illustrated the employment of a GAN-based simulated set of data that offers a controllable environment to train object detection models that achieve generalisation over diverse traffic environments [26].

Synthetic datasets and transfer learning methods also gained popularity. Synthetic datasets, which were developed with generative adversarial networks, allow researchers to mimic various traffic patterns to enhance model adaptability and accuracy. Transfer learning sees models learn from pre-trained weights from large databases to decrease training time and improve performance in other environments. A proposed example used SSD in conjunction with Kalman filters to track objects from one frame to the next, with effective object detection and tracking efficiency [27].

Real-time processing is still a primary need in intelligent transport systems applications and has contributed to the common use of hardware accelerators and optimisation deployment frameworks. TensorRT and ONNX tools have played a significant role in facilitating the implementation of real-time traffic light detection on edge devices. Optimisation frameworks maximise model inference to achieve speed-accuracy balancing. Used TensorRT to speed up YOLOv5-based systems to operate in real-time without degrading detection quality [15].

Finally, the use of three-dimensional localization methods is proving to be a promising area of development in traffic light detection. By pairing camera detections with high-definition maps and methods of depth estimation, developers are creating algorithms that can localize traffic signals with high accuracy in three dimensions. This is especially beneficial in autonomous cars that operate through complicated intersections where multiple traffic signals can be seen in a single image.

Today's trends in object detection in traffic light detection demonstrate a multidisciplinary perspective that unites research in deep learning, optimisation algorithms, and real-time processing technologies. Such innovations continue to open doors to more powerful, efficient, and flexible detection systems that overcome challenges in current traffic conditions and assure higher levels of accuracy of intelligent transport networks and autonomous vehicles, which help encounter the ever-persistent problems of illumination, occlusion, and cluttering.

TABLE II. Trends for Traffic Light Detection

TRENDS	DESCRIPTION
Anchor-Free Detection	Models like DARDet simplify tracking by predicting object locations without predefined bounding boxes, enhancing speed and accuracy.
Attention Mechanisms	Focus on key features in images to improve detection under occlusions and low-light conditions, as demonstrated in KCS-YOLO.
Multi-Camera Systems	Utilize data fusion from multiple cameras to improve robustness in complex scenarios, such as multi-intersection environments.
Hybrid Models	Combine deep learning (e.g., YOLO) with optimization algorithms (e.g., Firefly Algorithm) for better accuracy in dynamic conditions.
Synthetic Datasets	Use GAN-generated datasets to simulate diverse traffic scenarios, enhancing model generalization.
Transfer Learning	Leverage pre-trained weights to adapt models for new environments, reducing training time while maintaining effectiveness.
Real-Time Processing	Employ frameworks like TensorRT to optimize inference speed, enabling real-time tracking on edge devices.
3D Localization	Integrate depth estimation and high-definition maps for accurate traffic light positioning in three-dimensional space.

Table II illustrates on which way these trends are being implemented. It shows descriptively that these trends are usually used for encountering the problems that persist in the field of image processing of traffic light detection in intelligent

transportation systems, which are occlusion, cluttering, and illumination. The comprehensive result review of each trend for resolving the problems is in Section IV.

IV. Evaluation and Discussion

The radar chart in Fig. 1 illustrates the relative weaknesses and strengths of various trends of object detection specifically for traffic lights, providing a sense of how each trend performs according to important criteria, which include Accuracy, Speed, Robustness, Scalability, Simplicity, and Real-Time Processing when occurring with problems that persist in the field of image processing of traffic light detection in intelligent transportation systems which are occlusion, cluttering and illumination.

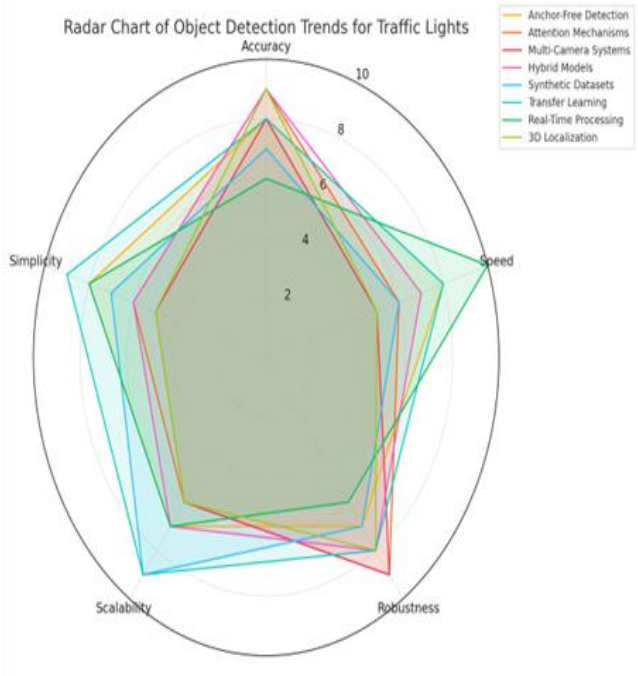


Fig. 1. Evaluation chart of object detection trends.

Anchor-free detection is quicker by eliminating predefined bounding boxes, but less robust under changing conditions. Attention mechanisms prioritize features better, offering balanced performance without excelling in any aspect. Multi-camera systems are spatially more aware and better handle occlusions, but are computationally more expensive. Hybrid models offer the best detection by merging various approaches, but could be challenging to tune. Synthetic datasets are more scalable and flexible but demand high-intensity processing. Transfer learning accelerates deployment with pre-trained models, but is hindered by drastic environment change. Real-time processing ensures fast detection but at the cost of stability in complex scenarios, and 3D localization improves detection accuracy with depth estimation but at the cost of computational efficiency.

It can be deduced from Fig. 2 that the effectiveness score derived from six evaluation criteria indicates strong performance on trends for "Multi-Camera Systems" and "Real-Time Processing", while the second-tier performance is "Anchor-Free Detection", "Hybrid Models", and "Transfer Learning". The bottom three which are "Attention Mechanisms", "3D Localization" and "Synthetic Datasets" gives the least optimum effectiveness score regarding object detection which in future can be a research field for improvement.

Fig. 2. Effectiveness of object tracking trends.

To ensure the probability of the charts, they are aligned with the deduction above. It is therefore appropriate to present the comparative data based on the previous studies with the comparative mean Average Precision (mAP) and Frame per second (FPS). The values presented are drawn from experimental results reported in previous peer-reviewed research papers, and actual performance may vary based on implementation and dataset. Table III shows evaluation metrics in terms of mAP and FPS.

TABLE III. Evaluation Metrics

APPROACH	EXAMPLE	mAP	FPS
Anchor-Free Detection	DARDet [30]	0.85	57.0
Attention Mechanisms	KCS-YOLO [22]	0.78	34.0
Multi-Camera Systems	aUToLights [15]	0.91	35.0
Hybrid Models	YOLOv4 + FA [25]	0.87	37.2
Synthetic Datasets	GAN-based dataset [33]	0.73	41.0
Transfer Learning	ResNet50 TL [28]	0.85	45.0
Real-Time Processing	YOLOv5 + TensorRT [4]	0.91	60.0
3D Localization	HD Maps + Depth Est. [34]	0.76	30.5

The comparative evaluation of trends (see Table III) is based on reported mean Average Precision (mAP) and Frames Per Second (FPS) from peer-reviewed studies, ensuring consistency across methods. For instance, real-time processing frameworks (e.g., YOLOv5 + TensorRT) achieve an average mAP of 0.91 with 60 FPS, indicating superior latency performance under occlusion. In contrast, hybrid models (YOLOv4 + FA) balance performance with mAP $\approx$ 0.87 and improved detection consistency under illumination variance.

It should be noted that the datasets referenced in these studies span diverse conditions, some synthetic (GAN-generated) and others real-world, including different regional signal encodings (e.g., Asia's five-color variant vs. Europe's three-light system) to ensure generalizability. These datasets cover day, night, rain, and fog scenarios, ensuring that each approach is evaluated under consistent environmental diversity. Thus, the comparative rankings are quantitatively grounded in published mAP and FPS benchmarks, representing relative trade-offs rather than subjective judgments.

No single approach is universally optimal; each performs best under specific conditions such as occlusion, illumination, or visual clutter. Anchor-free and real-time methods are better suited for edge deployment and latency-sensitive applications, such as on-device autonomous driving systems (occlusion). Meanwhile, attention-based and hybrid models offer greater robustness and accuracy, making them valuable in complex urban scenarios where precision is critical (occlusion, clutter, or illumination). Synthetic datasets and transfer learning provide scalability and domain adaptability, especially where labelled data is limited (occlusion). Multi-camera systems and 3D localization, while computationally intensive, can significantly enhance spatial awareness at intersections (occlusion, illumination, or clutter). In practice, a modular system that integrates several of these approaches, depending on environmental and application constraints, may offer the best balance of performance, accuracy, and cost-efficiency.

V. Conclusion

Traffic-light object detection plays an important role in intelligent transportation systems and autonomous driving, ensuring safe and smooth traffic flow. Traditional methods could not handle occlusions and illumination variations, leading to a shift toward hybrid models and deep learning. Multi-camera systems, 3D localization, and attention mechanisms are employed to boost detection accuracy, while transfer learning and synthetic datasets are utilized for better adaptability. Regional diversity and real-time implementation remain challenging. Future work must be directed at integrating various trends to develop scalable hybrid models for guaranteeing robust and efficient traffic light detection in continuously changing environments.

Future research should integrate multi-camera systems with real-time processing to achieve low-latency detection (target latency $\leq$ 20 ms) and employ hybrid models enhanced by transfer learning to reach higher precision (target mAP $\geq$ 0.95) in complex environments.

It is also important to distinguish how various environmental factors contribute to detection uncertainty. Occlusion primarily affects visibility consistency across frames, illumination alters color and contrast perception, while cluttering introduces background confusion. Addressing these through probabilistic modeling or uncertainty propagation frameworks (e.g., Bayesian CNNs) can further enhance model robustness.

Collectively, this review underscores the importance of combining complementary trends of hybrid learning, real-time frameworks, and multi-view data to build the next generation of resilient, adaptive traffic light detection systems for intelligent transportation networks.

Acknowledgment

The authors would like to thank the Ministry of Higher Education Malaysia (MOHE) for providing financial support under Fundamental Research Grant Scheme-Early Career Researcher (FRGS-EC) No. FRGS-EC/1/2024/ICT02/UMPSA/02/1 and Universiti Malaysia Pahang Al- Sultan Abdullah for laboratory facilities as well as additional financial support under Internal Research Grant RDU243705.

References

[1] S. Gautam and A. Kumar, "Image-based automatic traffic lights detection system for autonomous cars: a review," Multimed Tools Appl, vol. 82, no. 17, 2023, doi: 10.1007/s11042-023-14340-1

[2] S. S. Rao and S. R. Desai, "2 nd International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS 2021) Machine Learning based Traffic Light Detection and IR Sensor based Proximity Sensing for Autonomous Cars," 2021. [Online]. Available: https://ssrn.com/abstract=3883931

[3] J. M. De Guia and M. Deveraj, "Development of Traffic Light and Road Sign Detection and Recognition Using Deep Learning Towards Safe and Robust Sensor-Perception System of Autonomous Vehicle Development Research," 2024. [Online]. Available: www.ijacsa.thesai.org

[4] A. Phand, S. Bagade, N. Bandgar, and Prof. G. Wayal, "Real-Time Traffic Light Optimization Using AI and IOT," Int J Res Appl Sci Eng Technol, vol. 12, no. 4, pp. 3534–3539, Apr. 2024, doi: 10.22214/ijraset.2024.60686.

[5] S. Bali, T. Kumar, and S. S. Tyagi, "Development and performance evaluation of object and traffic light recognition model by way of deep learning," Indonesian Journal of Electrical Engineering and Computer Science, vol. 26, no. 3, 2022, doi: 10.11591/ijeecs.v26.i3.pp1486-1494.

[6] A. B. Zaidi and S. Zahera, "Real-time object detection and video monitoring in Drone System," International Research Journal of Engineering and Technology, 2023, [Online]. Available: www.irjet.net

[7] M. Z. Asy'ari, S. Filbert, and Z. L. Sukra, "Histogram of Oriented Gradients (HOG) and Haar Cascade with Convolutional Neural Network (CNN) Performance Comparison in the Application of Edge Home Security System," in Lecture Notes in Electrical Engineering, 2023. doi: 10.1007/978-3-031-29078-7_2.

[8] A. B. Amjoud and M. Amrouch, "Object Detection Using Deep Learning, CNNs and Vision Transformers: A Review," IEEE Access, vol. 11, 2023, doi: 10.1109/ACCESS.2023.3266093.

[9] S. Shah and J. Tembhurne, "Object detection using convolutional neural networks and transformer-based models: a review," Journal of Electrical Systems and Information Technology, vol. 10, no. 1, 2023, doi: 10.1186/s43067-023-00123-z.

[10] P. Shi, L. Yang, X. Dong, H. Qi, and A. Yang, "Research Progress on Multi-Modal Fusion Object Detection Algorithms for Autonomous Driving: A Review," Computers, Materials & Continua, vol. 0, no. 0, pp. 1–10, 2025, doi: 10.32604/cmc.2025.063205.

[11] M. Trigka and E. Dritsas, "A Comprehensive Survey of Machine Learning Techniques and Models for Object Detection," Sensors, vol. 25, no. 1, p. 214, Jan. 2025, doi: 10.3390/s25010214.

[12] R. Mohan, D. Cattaneo, F. Drews, and A. Valada, "Progressive Multi-Modal Fusion for Robust 3D Object Detection," Oct. 2024, [Online]. Available: http://arxiv.org/abs/2410.07475

[13] T. Zhao, B. Liu, Y. Gao, Y. Sun, M. Yuan, and X. Wei, "Rethinking Multi-modal Object Detection from the Perspective of Mono-Modality Feature Learning," Mar. 2025, [Online]. Available: http://arxiv.org/abs/2503.11780

[14] R. Padilla, S. L. Netto, and E. A. B. Da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," in International Conference on Systems, Signals, and Image Processing, 2020. doi: 10.1109/IWSSIP48289.2020.9145130.

[15] S. Wu, N. Amenta, J. Zhou, S. Papais, and J. Kelly, "aUToLights: A Robust Multi-Camera Traffic Light Detection and Tracking System," May 2023, doi: 10.1109/CRV60082.2023.00019.

[16] H. Tran Ngoc, K. Hoang Nguyen, H. Khanh Hua, H. Vu Nhu Nguyen, and L.-D. Quach, "Optimizing YOLO Performance for Traffic Light Detection and End-to-End Steering Control for Autonomous Vehicles in Gazebo-ROS2," 2023. [Online]. Available: www.ijacsa.thesai.org

[17] S. B. S. K. S. D. Prof. N. T. Vaibhav Bangadakar, "TRAFFIC RULE VIOLATION DETECTION WITH COMPUTER VISION," International Research Journal of Modernization in Engineering Technology and Science, May 2023, doi: 10.56726/irjmets40344.

[18] B. Suchithra, P. Sri, K. Chaitanya, and G. S. Kumar, "TRAFFIC LIGHT DETECTION AND CLASSIFICATION USING RESNET," JETIR, 2024. [Online]. Available: www.jetir.org

[19] V. Singh, "STM JOURNALS International Journal of Algorithms Design and Analysis Review Real-Time Object Detection and Tracking in Traffic Surveillance: Implementing Algorithms That Can Process Video Streams for Immediate Traffic Monitoring," 2025, doi: 10.37591/IJADAR.

[20] J. M. De Guia and M. Deveraj, "Development of Traffic Light and Road Sign Detection and Recognition Using Deep Learning Towards Safe and Robust Sensor-Perception System of Autonomous Vehicle Development Research," 2024. [Online]. Available: www.ijacsa.thesai.org

[21] M. A. Mohamad, H. Haron, and H. Hassan, "Flower Pollination Freeman Chain Code (FP-FCC) Extraction Algorithm for Handwritten Character Recognition," in New Trends in Intelligent Software Methodologies, Tools and Techniques, IOS Press, 2017, pp. 117–123. DOI: 10.3233/978-1-61499-800-6-117

[22] Q. Zhou, D. Zhang, H. Liu, and Y. He, "KCS-YOLO: An Improved Algorithm for Traffic Light Detection under Low Visibility Conditions," Machines, vol. 12, no. 8, Aug. 2024, doi: 10.3390/machines12080557.

[23] E. Trojovska, M. Dehghani, and P. Trojovsky, "Zebra Optimization Algorithm: A New Bio-Inspired Optimization Algorithm for Solving Optimization Algorithm," IEEE Access, vol. 10, 2022, doi: 10.1109/ACCESS.2022.3172789.

[24] J. M. De Guia and M. Deveraj, "Development of Traffic Light and Road Sign Detection and Recognition Using Deep Learning Towards Safe and Robust Sensor-Perception System of Autonomous Vehicle Development Research," 2024. [Online]. Available: www.ijacsa.thesai.org

[25] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," Apr. 2020, [Online]. Available: http://arxiv.org/abs/2004.10934

[26] H. Liu, Z. Tan, C. Tan, Y. Wei, J. Wang, and Y. Zhao, "Forgery-aware Adaptive Transformer for Generalizable Synthetic Image Detection."

[27] A. S. Talaat and S. El-Sappagh, "Enhanced aerial vehicle system techniques for detection and tracking in fog, sandstorm, and snow conditions," Journal of Supercomputing, vol. 79, no. 14, 2023, doi: 10.1007/s11227-023-05245-9.

[28] M. Nagendra Rao, T. Archana, M. Sirisha, T. Vishnuvardhan, and S. Shreya, "TRAFFIC LIGHT DETECTION AND CLASSIFICATION USING RESNET50."

[29] W. Chen and T. Shah, "Exploring Low-light Object Detection Techniques," Jul. 2021, [Online]. Available: http://arxiv.org/abs/2107.14382

[30] F. Zhang, X. Wang, S. Zhou, and Y. Wang, "DARDet: A Dense Anchor-free Rotated Object Detector in Aerial Images," Oct. 2021, doi: 10.1109/LGRS.2021.3122924.

[31] T. Ding, K. Feng, Y. Yan, Y. Wei, and T. Li, "An improved anchor-free method for traffic scene object detection," Multimed Tools Appl, vol. 82, no. 22, 2023, doi: 10.1007/s11042-023-15077-7.

[32] R. Soans and Y. Fukumizu, "Custom Anchorless Object Detection Model for 3D Synthetic Traffic Sign Board Dataset with Depth Estimation and Text Character Extraction," Applied Sciences (Switzerland), vol. 14, no. 14, Jul. 2024, doi: 10.3390/app14146352.

[33] X. Liu, A. Liu, J. L. Chen, and G. Li, "Impact of decomposition on time series bagging forecasting performance," Tour Manag, vol. 97, 2023, doi: 10.1016/j.tourman.2023.104725.

[34] T. Yin, X. Zhou, and P. Krähenbühl, "Center-based 3D Object Detection and Tracking," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2021. doi: 10.1109/CVPR46437.2021.01161.

[35] M. Rahman, K. Z. Zamli, and M. A. Mohamad, "A comparison of four metaheuristic algorithms for the problem of test redundancy reduction," in Proceedings of the 2023 12th International Conference on Software and Computer Applications (ICSCA), 2023, pp. 342–348. doi: 10.1145/3587828.3587879.

[36] M. A. Mohamad, M. A. Ahmad, and Z. Mustaffa, "Hybrid Honey Badger Algorithm with Artificial Neural Network (HBA-ANN) for Website Phishing Detection," Iraqi Journal for Computer Science and Mathematics, vol. 5, no. 3, pp. (671-682), 2024. doi: 10.52866/ijcsm.2024.05.03.041

Conversational AI-Powered VR Development Model for Tourism Promotion in Thailand: Expert Assessment and Stakeholder Acceptance

Jenasama Srihirun[1], Kridsanapong Lertbumroongchai[2], Vitsanu Nittayathammakul[3], Pimon Kaewdang[4]
Department of Tourism and Hospitality Management-Faculty of Liberal Arts, Rajamangala University of Technology Thanyaburi, Pathum Thani, Thailand[1, 4]
Division of Computer Science-Faculty of Science and Technology,
Rajamangala University of Technology Thanyaburi, Pathum Thani, Thailand[2]
Ratchasuda Institute-Faculty of Medicine Ramathibodi Hospital, Mahidol University, Nakhon Pathom, Thailand[3]

***Abstract*—Thailand's tourism sector increasingly requires immersive digital innovations that preserve local identity while enhancing visitor engagement. However, there remains a lack of a comprehensive model to guide such developments. This study aims to propose the Conversational AI-powered Virtual Reality Development Model for Tourism Promotion in Thailand, providing an integrated and context-specific framework suitable for practical implementation. A Design and Development Research (DDR) methodology (Type II) was employed in three stages: 1) synthesizing essential components through a scoping review, 2) constructing and validating the model via expert panels using the Content Validity Index (CVI) analysis, and 3) assessing suitability and acceptance through expert evaluation and stakeholder surveys. The model developed in this study, referred to as the 4Ds Model, contributes new knowledge by integrating conversational AI and virtual reality within a four-phase structure—Discover, Design, Develop, and Deploy—supported by five enabling capitals: human, cultural, technological, informational, and financial. The Deploy phase modifies the AISAS communication framework into AICAS (Attention, Interest, Chat, Action, Share) to illustrate the function of conversational AI in improving user interaction and engagement within the context of tourism in Thailand. Results indicated high expert ratings of suitability and strong stakeholder intention to adopt. Multiple regression analysis revealed that technological self-efficacy, perceived interactivity, and perceived tourism benefits were significant predictors, explaining 73.3% of the variance in behavioral intention. The findings demonstrate both the theoretical advancement in AI–VR integration and the practical readiness of the 4Ds Model as a culturally aligned roadmap for digital tourism transformation in Thailand.**

Keywords*—Virtual reality; conversational AI; Model Development; digital tourism; technology adoption

I. Introduction

Tourism is a major economic sector worldwide, playing a vital role in job creation, income generation, and cultural exchange. In recent years, the industry has faced significant challenges, especially due to the COVID-19 pandemic, which accelerated the shift toward digital engagement and virtual-tourism solutions [1]. As a response, technologies such as virtual reality (VR) and conversational artificial intelligence (CAI) have gained prominence as innovative tools for enhancing tourism marketing and traveler engagement [2].

VR allows users to experience simulated environments that replicate real-world tourist destinations. These immersive environments can shape destination image, emotional attachment, and behavioral intention among potential tourists [3], [4]. Multisensory VR—which integrates sound, vision, and other sensory inputs—has been shown to increase emotional engagement and perceived presence, particularly when personalized to specific user groups [4].

CAI plays a complementary role by enabling real-time dialogue and personalized assistance in virtual environments. Services such as language translation, itinerary recommendations, and interactive storytelling help reduce uncertainty and enhance satisfaction for users navigating virtual-tourism experiences [5]. Despite growing global applications of VR and AI in tourism, their integrated use—particularly within culturally diverse and linguistically unique regions like Thailand—remains underexplored [6].

However, existing studies rarely provide a structured model that systematically integrates VR and CAI for tourism promotion, leaving a methodological gap in both research and practice. Without such a structured model, integration efforts risk remaining fragmented, thereby limiting scalability, policy alignment, and long-term impact. For Thailand specifically, aligning immersive technologies with national tourism strategies and embedding local cultural identity are critical considerations, yet current models offer little structured guidance for achieving such alignment.

Research consistently highlights the importance of interactivity, presence, and satisfaction in shaping tourist behavior in VR environments. When VR is combined with conversational systems, emotional immersion is heightened and human-like engagement is created, effectively simulating tour guides or hosts to enrich the tourist's pre-visit experience [4], [7].

This study proposes the Conversational AI-Powered VR Development Model for Tourism Promotion in Thailand as an innovative response to the identified research gap. The model combines immersive virtual environments with conversational

interfaces to enhance engagement and personalization, while remaining aligned with digital tourism strategies. To systematically examine the model, the study focuses on three key aspects: 1) synthesizing its core components, 2) validating its theoretical soundness, and 3) evaluating its suitability and acceptability from expert and stakeholder perspectives. In this study, stakeholders refer not to tourists but to public-sector tourism officers, destination management professionals, and practitioners in provincial tourism promotion units, whose perspectives are essential for policy alignment and implementation. The research also seeks to identify the key factors that predict stakeholders' behavioral intention to adopt the model. These aims directly inform the research objectives and questions outlined in the following section.

II. Research Objectives

1) To synthesize the key components of the conversational AI-powered VR development model for tourism promotion in Thailand.

2) To develop and validate the conversational AI-powered VR development model in terms of theoretical soundness.

3) To evaluate the model's suitability as perceived by domain experts in the tourism sector.

4) To evaluate the model's acceptability as perceived by stakeholders and to identify the key factors predicting their behavioral intention to use the model.

III. Research Questions

1) RQ1: What are the key components of the conversational AI-powered VR development model for tourism promotion in Thailand?

2) RQ2: To what extent is the synthesized model valid in terms of theoretical soundness?

3) RQ3: To what extent do domain experts in the tourism sector perceive the proposed model as suitable?

4) RQ4: To what extent do stakeholders in the tourism sector perceive the proposed model as acceptable, and which overall and specific factors significantly predict their behavioral intention to use it?

IV. Literature Review

A. Virtual Reality in Tourism Promotion

Virtual reality (VR) has been widely adopted as a technology-based marketing tool in the tourism industry because it can simulate real-world experiences. Research indicates that VR enables travellers to pre-experience destinations, which in turn shapes their emotional engagement and travel intentions [8]. The immersive quality of VR—particularly in multisensory formats—enhances presence and enjoyment [9]. Studies further suggest that VR can function as a substitute for real visits to sensitive or remote locations, supporting both destination marketing and conservation goals [10]. Recent reviews emphasise VR's growing relevance in the post-pandemic era, as the technology-maintained destination visibility during periods of restricted travel [11]. Moreover, VR has proven effective in reshaping traveller behaviour and expectations, increasing the potential for its long-term use in destination planning [12].

B. Conversational AI in Tourism Applications

Conversational artificial intelligence (CAI)—including chatbots and virtual assistants—has reshaped tourist services by providing real-time, personalized, and multilingual communication. In tourism and hospitality settings, conversational agents improve user satisfaction, assist with bookings, and reduce staff workload [13]. These systems leverage natural language processing to recognize user intent and adjust responses contextually, creating more natural interactions. Tourists also report positive emotional responses to AI-powered services, especially when these systems are integrated seamlessly across digital platforms. Trust, perceived usefulness, and enjoyment are key factors that drive the continued adoption of AI in tourism [13]. Nevertheless, most existing studies examine AI in isolation, focusing on functional or transactional roles with limited attention to its immersive and experiential potential—particularly in tandem with VR.

C. Integrated Use of VR and Conversational AI

Although VR and CAI have each been applied in tourism, research on their integrated use is still emerging. A combined system offers both immersive visual engagement and interactive verbal communication, effectively simulating local guides or cultural narrators within virtual environments. For instance, integrating speech recognition and natural-language processing into VR environments has been shown to improve user experience and language accessibility, especially for tourists with limited English proficiency [14]. Such systems can adaptively deliver information, answer queries, and provide decision support, making virtual tourism more personalised and engaging. However, there remains a lack of formalized models detailing how these integrated systems should be developed, validated, and deployed—particularly in alignment with national tourism strategies.

D. Cultural and Linguistic Personalization in Tourism

Cultural and linguistic personalization is critical in tourism-technology design, especially in diverse destinations such as Thailand. AI systems that support native-language interactions and culturally relevant content strengthen trust and reduce uncertainty among tourists [13]. In VR environments, weaving culturally nuanced storytelling and symbolic elements—such as temple architecture or local customs—deepens emotional resonance and visitor satisfaction [8]. Recent studies also indicate that tourists perceive greater authenticity and enjoyment when virtual systems reflect their cultural context and accommodate preferred languages [14]. Nevertheless, few development models systematically incorporate cultural and linguistic personalization—particularly those aligned with public-sector objectives for national tourism promotion.

E. Determinants of Technology Adoption in Tourism

The adoption of the proposed Conversational AI-Powered VR Development Model by stakeholders—defined in this study as public-sector tourism officers, destination management professionals, and practitioners in provincial tourism promotion units—depends on several interrelated psychological, organizational, and technological factors. These determinants are consistent with established adoption frameworks such as the Technology Acceptance Model (TAM) [15] and the Unified Theory of Acceptance and Use of Technology (UTAUT) [16],

which highlight perceived usefulness, ease of use, facilitating conditions, and behavioral intention as critical predictors of technology adoption.

One of the most critical factors is technological self-efficacy, referring to stakeholders' confidence in their ability to understand, operate, and apply new technologies in their work. Prior studies confirm that self-efficacy strongly influences the adoption of digital platforms in tourism and hospitality [17], [18]. High self-efficacy supports perceived ease of use and perceived usefulness—core constructs in TAM—and shapes favorable attitudes toward adopting innovative tools [19].

Another determinant is attitude toward tourism technology. Stakeholders who perceive VR and AI as valuable for enhancing tourism experiences and promoting destinations are more willing to support and invest in such tools. This aligns with TAM's emphasis on attitude as a mediator of adoption decisions [20] and with UTAUT's construct of performance expectancy [16].

Facilitating conditions—such as institutional support, infrastructure, budget allocation, and training—either enable or constrain system adoption. As emphasized in UTAUT, strong facilitating conditions increase the likelihood of adoption, especially when coupled with organizational readiness [16], [21].

Perceived accessibility—including ease of use, availability across devices, and language inclusivity—further influences adoption. Systems that minimize technical barriers and support local languages are particularly attractive to public-sector tourism personnel in linguistically diverse contexts [14].

Perceived interactivity is a decisive factor for technologies designed for tourism promotion. Dynamic, real-time engagement (e.g., conversational agents embedded in VR) enhances the perceived quality of communication experiences, encouraging adoption [14].

Finally, perceived tourism benefits reflect beliefs that the system can improve destination image, educate tourists, or stimulate visitation. When stakeholders perceive strong benefits aligned with policy priorities, they are more likely to adopt and champion the system [17].

These determinants—self-efficacy, attitudes, facilitating conditions, accessibility, interactivity, and perceived benefits—reflect validated constructs in TAM and UTAUT while addressing the specific needs of Thailand's public-sector tourism stakeholders. Unlike prior studies that examined these factors in isolation or emphasized tourists as the primary users, this study integrates them into a structured model. The determinants are empirically tested as predictors of stakeholders' perceived acceptability of the model and their behavioral intention to use it, underscoring the study's contribution in bridging technology acceptance theory with the policy and operational realities of tourism promotion in culturally diverse contexts.

V. Methodology

This study employed a Design and Development Research (DDR) Type II methodology, which is well-suited for synthesizing, developing, and evaluating innovative models in real-world contexts where no established models exist. According to Richey and Klein (2014), DDR Type II emphasizes both theoretical rigor and practical applicability. The research proceeded in three stages: 1) Model Synthesis, 2) Model Development and Validation, and 3) Model Evaluation.

A. Stage I: Model Synthesis

In the first stage, a scoping review was conducted to identify and synthesize the key components of the conversational AI-powered VR development model for tourism promotion in Thailand. Following the framework of Arksey and O'Malley [23], searches were performed in Scopus, IEEE Xplore, and ThaiJO databases covering publications between 2019 and 2025. The search strategy combined keywords such as "virtual reality", "conversational AI", "smart tourism", "digital tourism", "cultural communication", and "government tourism innovation". Three instruments supported this stage: a) Study Screening and Eligibility Checklist, b) Data Charting Form, and c) Evidence Synthesis Matrix. The results of this stage provided the preliminary components of the development model.

B. Stage II: Model Development and Validation

The second stage refined and validated the preliminary model. A focus group discussion (FGD) was organized with nine purposively selected experts specializing in tourism technology, VR/AI systems, and Thai cultural media. The FGD followed a semi-structured protocol described in [24]. Insights collected from the discussion were thematically analyzed to enhance the structure, terminology, and contextual alignment of the model. Subsequently, the revised model underwent content validation using the Content Validity Index (CVI). A separate panel of five experts—holding doctoral degrees and with at least five years of relevant experience—rated each component on a four-point relevance scale. All item-level CVIs (I-CVI) exceeded 0.80, and the scale-level CVI/Average (S-CVI/Ave) was 0.94, surpassing the 0.90 threshold recommended in [25]. These results confirmed strong content validity.

C. Stage III: Model Evaluation

This stage evaluated the suitability and acceptability of the proposed model through expert and stakeholder feedback, ensuring both theoretical soundness and practical applicability.

1) Participants: This stage engaged two categories of participants to capture complementary perspectives. The expert category included five purposively selected individuals, each holding a doctoral degree in a relevant discipline, with at least five years of professional or research experience in tourism development, immersive technology, or cultural communication, and recognized contributions such as publications, funded projects, or leadership roles. The stakeholder category comprised 120 participants recruited through a multi-stage sampling process. First, Thailand's four geographical regions—North, Central and East, Northeast, and South—were identified. Next, provincial tourism promotion units within each region were targeted. Finally, thirty participants from each region, including government officials, administrators, and practitioners directly engaged in tourism promotion, were selected.

2) Instruments: Separate instruments were employed for experts and stakeholders, each tailored to its respondent group. Both used a five-point Likert scale (1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly Agree). The expert evaluation form consisted of twelve items corresponding to the four phases of the model—Discover, Design, Develop, and Deploy—focusing on practicality, clarity, and policy alignment. The stakeholder survey was divided into two parts: the first part collected demographic and contextual information, and the second part measured six independent variables, namely Technological Self-Efficacy (X_1), Attitude towards Technology in Tourism (X_2), Facilitating Conditions (X_3), Perceived Accessibility (X_4), Perceived Interactivity (X_5), and Perceived Tourism Benefits (X_6), as well as one dependent variable, Behavioral Intention to Use the Model (Y). These constructs were adapted from previously validated studies [15–19] and contextualized for Thai tourism. Both instruments were pilot-tested with thirty respondents, and their quality was confirmed through expert review (CVI ≥ 0.90) and reliability analysis (Cronbach's alpha ≥ 0.80) [25].

3) Data collection procedures: Experts were contacted individually and asked to complete the evaluation form within two weeks. Stakeholders were surveyed in collaboration with regional tourism offices and networks, using both online and in-person administration depending on accessibility. Data collection lasted for one month. Informed consent was obtained from all participants, and confidentiality as well as voluntary participation were assured.

4) Data analysis: Data from the expert group (n = 5) were analyzed using descriptive statistics, including means and standard deviations, combined with qualitative synthesis. Inferential analysis was avoided due to the small sample size. Data from the stakeholder group (n = 120) were analyzed using both descriptive and inferential statistics. Descriptive statistics were applied to summarize responses, Pearson's correlation was used to examine relationships among variables, and Multiple Regression Analysis (MRA) was performed to identify predictors of behavioral intention (Y) from the six independent variables (X_1–X_6). Prior to regression, assumptions were tested for linearity, independence of errors, normality, homoscedasticity, and multicollinearity, with thresholds set at VIF < 10 and Tolerance > 0.1. Sample size adequacy was established based on Green's rule of 15–20 cases per predictor [26], confirming that the sample of 120 participants was sufficient.

5) Ethical considerations: This study adhered to the Declaration of Helsinki, the Belmont Report, the CIOMS Guidelines, and the ICH-GCP framework. Ethical approval was obtained from the Human Research Ethics Committee of Rajamangala University of Technology Thanyaburi, Thailand (COA No. 73, RMUTT_REC No. Exp 73/68) through an expedited review on July 7, 2025. All participants provided written informed consent, with full rights to withdraw at any stage.

VI. Results

A. The Key Components of the Conversational AI-Powered VR Development Model

The conceptual foundation of the Conversational AI-powered VR Development Model for tourism promotion in Thailand was derived from a synthesis of international and local scholarship in virtual tourism, artificial intelligence, human–computer interaction, and cultural experience design. Based on this review, the model is structured around three interrelated domains—Virtual Reality, Conversational AI, and Human-Centered Design—supported by five enabling capitals that ensure sustainable implementation within Thailand's tourism ecosystem.

Virtual Reality (VR) functions as the primary medium for immersive destination simulation. The literature identifies three levels of immersion—non-immersive, semi-immersive, and fully immersive—with the latter two proving most effective in strengthening spatial presence and influencing travel intention [27–29]. Core components include 3D environmental modelling [52], the integration of cultural content, and multisensory feedback systems [30–32]. Scenario design plays a critical role in tourism applications, where meaningful storytelling enhances emotional engagement with cultural landmarks and traditions [33].

Conversational AI adds an intelligent, interactive layer that facilitates real-time dialogue through natural language processing (NLP), automatic speech recognition, and context-aware dialogue management [34], [35]. These capabilities enable virtual agents to act as culturally aligned local guides, providing personalised and linguistically appropriate recommendations [36], [37]. Research highlights that conversational flow should reflect national etiquette, local expressions, and culturally rooted storytelling in order to resonate with diverse audiences [38], [39].

The human-centered design domain focuses on user engagement, perception, and emotional response. This includes modelling diverse tourist personas, designing accessible interfaces, and applying affective strategies to sustain attention [40], [41]. Emphasis on privacy, inclusivity, and cultural accuracy is essential to building trust and ensuring respectful experiences [42], [43]. In the Thai context, systems must not only deliver visual appeal but also convey the symbolic and cultural depth of places, people, and traditions.

Supporting these domains are five enabling capitals identified in tourism innovation literature: human capital (skills and expertise), cultural capital (traditions and narratives), technological capital (infrastructure and platforms), information capital (user data and content libraries), and financial capital (public and private investment) [44–46]. Together, these resources form the foundation for sustainable development, scalability, and policy alignment in Thailand's digital tourism initiatives.

B. The Proposed Model for Tourism Promotion in Thailand

The Conversational AI-powered VR Development Model, hereafter referred to as the 4Ds Model, was developed and validated during Stage II of this study. It comprises four

structured phases—Discover, Design, Develop, and Deploy—as illustrated in Fig. 1. The model translates insights from the synthesis of Stage I into a practical framework, refined through expert input to ensure theoretical grounding, contextual relevance, and alignment with Thailand's public-sector tourism promotion priorities.

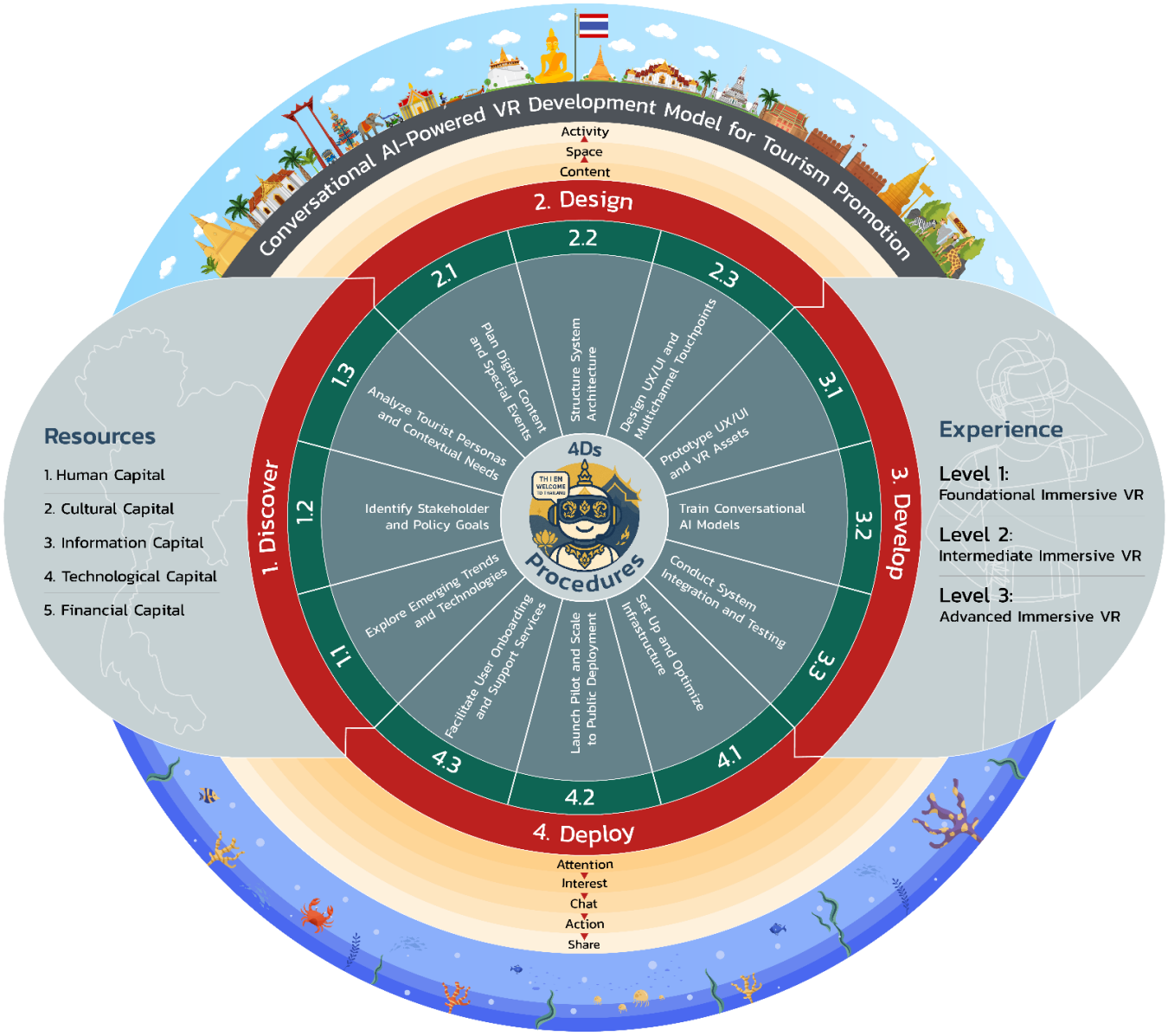


Fig. 1. The conversational AI-powered VR development model for tourism promotion in Thailand.

The Conversational AI-powered VR Development Model for Tourism Promotion in Thailand is structured into four principal phases, comprising a total of twelve systematically organized sub-steps, as follows:

1) Discover (D1): This phase focuses on strategic and contextual exploration to ensure that immersive technology development aligns with stakeholder priorities and destination-specific needs. It involves considering the availability and readiness of five enabling capitals—human, cultural, technological, informational, and financial—while also identifying target user personas and tourism communication goals. This ensures that subsequent phases are grounded in real-world constraints and opportunities. It comprises three sub-steps:

a) Explore emerging trends and technologies: To investigate recent developments in virtual reality, conversational AI, and digital tourism innovations.

b) Identify stakeholder and policy goals: To review national/local tourism policies and align digital initiatives with governmental priorities.

c) Analyze tourist personas and contextual needs: To understand target tourist segments, their motivations, behaviors, and cultural expectations.

2) Design (D2): This phase emphasizes conceptual planning and experience architecture across three core dimensions: content, space, and activity. It translates insights from the Discover phase into an actionable blueprint, detailing narrative structure, spatial navigation, and interactive elements that reflect cultural heritage. The design ensures that immersive tourism experiences are coherent, contextually meaningful, and engaging. It comprises three sub-steps:

a) Plan digital content and special events: To design thematic narratives and cultural programming across physical, digital, and hybrid formats.

b) Structure system architecture: To define the platform's modules, integration flow, and interaction logic.

c) Design UX/UI and Multichannel Touchpoints: To develop user interfaces and engagement pathways across devices and formats.

3) Develop (D3): This phase involves the technical realization of the system, including asset creation, system integration, and iterative testing. It operationalizes three progressive levels of immersive experience—foundational, intermediate, and advanced—each tailored to different user expectations and technological capabilities. It also includes training conversational AI agents to support real-time user interaction. It comprises three sub-steps:

a) Prototype UX/UI and VR assets: To produce visual, spatial, and interactive elements reflecting cultural context.

b) Train conversational AI models: To develop and localize intelligent agents capable of multilingual, context-aware interaction, ensuring that users from diverse linguistic backgrounds—such as Thai, English, Chinese, and regional ASEAN languages—can access tourism content seamlessly. This involves training models not only to handle literal translation but also to adapt cultural nuances, idiomatic expressions, and domain-specific terminology, thereby supporting inclusive, natural, and contextually appropriate communication across different audiences.

c) Conduct system integration and testing: To combine all components into a working prototype and evaluate its functionality and usability.

4) Deploy (D4): This final phase focuses on implementation, public rollout, and ongoing engagement. It adopts the AICAS strategy—Attention, Interest, Chat, Action, and Share—adapted from the AISAS model (Attention, Interest, Search, Action, Share) [47], [48]. In this adaptation, Search is replaced with Chat to reflect dialogic, real-time inquiry within conversational AI–enabled VR environments, where users obtain information through natural-language interaction rather than traditional keyword search. This modification aligns the engagement funnel with CAI-led user journeys while preserving the intent of the original AISAS framework. It comprises three sub-steps:

a) Set up and optimize infrastructure: To ensure technical readiness and system reliability for public deployment.

b) Launch pilot and scale to public deployment: To initiate small-scale trials and refine before wide release.

c) Facilitate user onboarding and support services: To provide orientation, assistance, and ongoing user engagement support.

C. The Suitability of the Proposed Model from Expert Perspectives

Five experts were invited to evaluate the overall suitability of the proposed 4Ds Model for tourism promotion in Thailand, focusing on its twelve sub-steps across the four phases. Their assessments reflected perceptions of the clarity, relevance, and appropriateness of each sub-step within the Thai tourism context. As shown in Table I, mean scores ranged from 4.40 to 4.80, with an overall mean of 4.55. Although these results are based on a small expert panel (n = 5) and should therefore be interpreted as indicative rather than generalizable, the standard deviations remained consistent at 0.53, suggesting strong agreement among the experts.

TABLE I. THE SUITABILITY OF THE CONVERSATIONAL AI-POWERED VR DEVELOPMENT MODEL FOR TOURISM PROMOTION

Assessment Items		M	SD	Suitability Level
Phase	**Sub-steps**			
1. Discover (D1)	1.1 Explore Emerging Trends and Technologies	4.40	0.55	Agree
	1.2 Identify Stakeholder and Policy Goals	4.60	0.55	Strongly Agree
	1.3 Analyze Tourist Personas and Contextual Needs	4.40	0.55	Agree
2. Design (D2)	2.1 Plan Digital Content and Special Events	4.60	0.55	Strongly Agree
	2.2 Structure System Architecture	4.60	0.55	Strongly Agree
	2.3 Design UX/UI and Multichannel Touchpoints	4.40	0.55	Agree
3. Develop (D3)	3.1 Prototype UX/UI and VR Assets	4.40	0.55	Agree
	3.2 Train Conversational AI Models	4.80	0.45	Strongly Agree
	3.3 Conduct System Integration and Testing	4.60	0.55	Strongly Agree
4. Deploy (D4)	4.1 Set Up and Optimize Infrastructure	4.40	0.55	Agree
	4.2 Launch Pilot and Scale to Public Deployment	4.80	0.45	Strongly Agree
	4.3 Facilitate User Onboarding and Support Services	4.60	0.55	Strongly Agree
Overall		4.55	0.53	Strongly Agree

Qualitative feedback reinforced these quantitative results. Experts particularly highlighted the Design phase, which they valued for its structured articulation of content modules, spatial navigation, and interaction logic, and the Deploy phase, which was praised for operationalizing the AICAS strategy (Attention, Interest, Chat, Action, Share) that reflects dialogic, real-time engagement in CAI–VR environments. Collectively, these judgments indicate that the 4Ds Model is conceptually robust, technically coherent, and contextually applicable to Thailand's tourism promotion efforts. Experts further emphasized that the model's flexible procedures make it adaptable across six categories of tourist attractions: 1) wellness and lifestyle tourism (e.g., spas, retreats, health resorts); 2) cultural and heritage tourism (e.g., temples, historical parks, historical museums); 3) nature and ecotourism (e.g., national parks, marine reserves); 4) creative and community-based tourism (e.g., local crafts, village experiences); 5) educational or edutainment tourism (e.g., zoos, aquariums, science centers); and 6) academic and medical tourism (e.g., universities, teaching hospitals, specialty clinics).

D. The Descriptive Statistics and Correlation Analysis of Stakeholders' Responses

A total of 120 stakeholders participated in the survey, comprising 73 females (60.83%) and 47 males (39.17%). Their ages ranged from 18 to 63 years, with a mean of 28.15 years (SD = 10.96) and a median of 24 years, indicating that the majority were young to middle-aged adults. In terms of educational attainment, most respondents held a bachelor's degree (59.17%), followed by those with a master's degree (30.83%) and a doctoral degree (10.00%).

With respect to occupation, the largest group consisted of public-sector tourism officers and staff in provincial tourism promotion units (45.00%). They were followed by private-sector tourism professionals such as tour guides, hotel staff, and travel agency operators (35.00%), and by media professionals or content creators specializing in tourism promotion (20.00%). This composition reflects a balanced mix of public officials, private-sector practitioners, and communication agents, consistent with the study's emphasis on stakeholder perspectives rather than tourists.

Regarding prior experience with technology in tourism, most respondents had never used VR (40.83%), while others had used it once (26.67%), two to three times (20.83%), or four to five times and above (11.67%). In contrast, AI applications for tourism were more widely adopted: 39 respondents (32.50%) had never used them, 28 (23.33%) had used them once, 31 (25.83%) had used them two to three times, and 22 (18.33%) had used them four to five times or more.

The descriptive statistics and Pearson's correlation coefficients of the stakeholder survey responses are presented in Table II. The results show that the mean values of the independent variables ranged from 4.19 to 4.47, while the dependent variable— Behavioral Intention to Use the Conversational AI-powered VR Development Model — had a mean score of 4.33 (SD = 0.65). As all means were above the scale midpoint (3), this indicates a consistently high level of acceptance toward the proposed model among stakeholders. Furthermore, all independent variables were positively and significantly correlated with behavioral intention (Y). The strongest correlations with Y were observed for Perceived Interactivity (r = .806), Perceived Tourism Benefits (r = .800), and Perceived Accessibility (r = .721). Although some intercorrelations were relatively high (e.g., X_5 and X_6), none exceeded problematic thresholds, suggesting that multicollinearity was not severe at this stage. It should also be noted that correlation analysis does not establish causality but provides a statistical foundation for subsequent regression testing.

TABLE II. THE DESCRIPTIVE STATISTICS AND CORRELATION MATRIX

Variable	M	SD	X1	X2	X3	X4	X5	X6	Y
1. Technological Self-Efficacy (X_1)	4.20	0.54	1						
2. Attitude towards Technology in Tourism (X_2)	4.47	0.47	.500	1					
3. Facilitating Conditions (X_3)	4.19	0.64	.596	.438	1				
4. Perceived Accessibility (X_4)	4.28	0.60	.534	.546	.741	1			
5. Perceived Interactivity (X_5)	4.28	0.57	.591	.537	.785	.802	1		
6. Perceived Tourism Benefits (X_6)	4.32	0.54	.517	.510	.675	.741	.821	1	
7. Behavioral Intention to Use the Conversational AI-powered VR Development Model (Y)	4.33	0.65	.602	.509	.703	.721	.806	.800	1

E. Assumption Testing for Multiple Regression

Regression assumptions were tested prior to model estimation. The Durbin–Watson statistic was 1.70, confirming independence of errors. The Shapiro–Wilk test ($p < .001$) indicated some deviation from perfect normality; however, given the adequate sample size (n = 120), residuals were approximately normally distributed, as supported by histogram and Q–Q plots. This meets the robustness criteria for regression with moderate to large samples [49], [50].

Homoscedasticity was supported by residual plots showing no funnel pattern. Variance Inflation Factor (VIF) values ranged between 1.61 and 5.05, below the threshold of 10, indicating no serious multicollinearity, though X_5 showed relatively higher collinearity. Overall, the assumptions were sufficiently met.

F. Multiple Regression Results

As presented in Table III, the multiple regression model significantly predicted stakeholders' behavioral intention to use the Conversational AI-powered VR Development Model, $F(6, 113) = 51.689$, $p < .001$. The six predictors together explained 73.3% of the variance in behavioral intention ($R^2 = .733$), with an adjusted R^2 of .719, indicating strong explanatory power of the model within this dataset.

TABLE III. MODEL SUMMARY OF MULTIPLE REGRESSION

R	R^2	Adjusted R^2	F	Sig. (p)
.856	.733	.719	51.689	< .001

TABLE IV. REGRESSION COEFFICIENTS OF PREDICTORS ON BEHAVIORAL INTENTION TO USE THE CONVERSATIONAL AI-POWERED VR DEVELOPMENT MODEL

Predictor	B	SE	Beta	t	Sig.	95% CI for B	VIF
Constant	-0.507	0.33	–	-1.54	.129	[-1.16, 0.15]	–
1. Technological Self-Efficacy (X_1)	0.174	0.077	0.146	2.25	.027*	[0.02, 0.33]	1.78
2. Attitude towards Technology in Tourism (X_2)	0.029	0.085	0.021	0.34	.736	[-0.14, 0.20]	1.61
3. Facilitating Conditions (X_3)	0.085	0.087	0.084	0.98	.328	[-0.09, 0.26]	3.07
4. Perceived Accessibility (X_4)	0.068	0.097	0.063	0.70	.484	[-0.12, 0.25]	3.40
5. Perceived Interactivity (X_5)	0.313	0.124	0.277	2.53	.013*	[0.07, 0.55]	5.05
6. Perceived Tourism Benefits (X_6)	0.461	0.106	0.384	4.36	.000***	[0.25, 0.67]	3.29

Note: *p < .05, ***p < .001

As shown in Table IV, three variables emerged as significant predictors: Technological Self-Efficacy (X_1), Perceived Interactivity (X_5), and Perceived Tourism Benefits (X_6). The remaining predictors—Attitude towards Technology in Tourism (X_2), Facilitating Conditions (X_3), and Perceived Accessibility (X_4)—did not reach statistical significance.

The regression equation is presented in Eq. (1):

$$Y = 0.507 + 0.174X_1 + 0.029X_2 + 0.085X_3 + 0.068X_4 + 0.313X_5 + 0.461X_6 \quad (1)$$

where, Y represents behavioral intention, X_1 denotes technological self-efficacy, X_2 denotes attitude towards technology in tourism, X_3 denotes facilitating conditions, X_4 denotes perceived accessibility, X_5 denotes perceived interactivity, and X_6 denotes perceived tourism benefits.

A simplified model with only significant predictors is shown in Eq. (2):

$$Y = 0.507 + 0.174X_1 + 0.313X_5 + 0.461X_6 \quad (2)$$

From the stakeholder perspective, multiple regression analysis indicated that X_1 (technological self-efficacy), X_5 (perceived interactivity), and X_6 (perceived tourism benefits) were significant predictors of Y (behavioral intention). Collectively, these three factors accounted for 73.3% of the variance in behavioral intention, demonstrating the model's strong explanatory power and practical relevance for tourism promotion in Thailand. However, the findings should be interpreted as context-specific evidence drawn from the surveyed stakeholder group—public-sector tourism officers, destination management professionals, and provincial tourism practitioners—and therefore may not be directly generalizable to tourists or to broader populations beyond the study sample.

VII. DISCUSSION

This study set out to design, validate, and evaluate the Conversational AI-powered VR Development Model for tourism promotion. The results provided convergent evidence across expert evaluations and stakeholder responses, confirming both theoretical soundness and practical feasibility. In the following discussion, the findings are organized around the four research questions (RQ1 to RQ4) and further interpreted in relation to existing theories and empirical studies in VR, AI, and technology adoption.

Addressing RQ1, which investigated the key components of the model, the study confirmed that the Conversational AI-powered VR Development Model can be systematically structured around three interrelated domains—Virtual Reality, Conversational AI, and Human-Centered Design—supported by five enabling capitals: human, cultural, technological, informational, and financial. This configuration echoes the systematic reviews of Calisto and Sarkar [11] and Mariani et al. [34], who highlighted the need for integrated frameworks that unify immersive technologies with user-centered design principles, and resonates with broader system design methodologies that emphasize iterative cycles of requirements analysis, architectural specification, and prototyping [22], [37]. A distinctive feature of the model is its explicit articulation of three progressive levels of immersion—non-immersive, semi-immersive, and fully immersive—corresponding respectively to foundational, intermediate, and advanced experiences. This dual framing bridges technical VR typologies with operational deployment categories, aligning with Slater and Sanchez-Vives [28] and with Anaya-Sánchez et al. [51], who empirically demonstrated that immersion moderates the impact of VR experiences on destination image and visit intentions. Furthermore, the model structures immersive experiences across content, space, and activity dimensions, offering developers a blueprint that is both flexible and technically rigorous, ensuring applicability across diverse organizational and user contexts [15], [16].

In relation to RQ2, which examined theoretical validity, both content validity testing and expert review indicated strong soundness (I-CVI≥ .80; S-CVI/Ave = .94), following guidelines by Yusoff [25]. Experts noted that the phased, modular structure is consistent with Thailand's policy directions for technology-enhanced tourism promotion. This contextual applicability reflects Wu and Zhang [39], who emphasized that immersive technologies must reflect policy and cultural conditions to achieve sustainable adoption. It also resonates with Guttentag [10] and Phoong et al. [29], who stressed that VR initiatives must consider real-world constraints and tourism development trajectories to evolve from novelty to mainstream adoption. By embedding system design within both technological and policy

contexts, the 4Ds Model demonstrates theoretical rigor while preserving contextual relevance.

For RQ3, concerning suitability, experts rated the model highly across phases (M = 4.55, SD = 0.53). The Design phase was valued for its structured articulation of content modules, spatial navigation, and interaction logic, consistent with UX principles in VR system design [37], [42]. The Deploy phase was praised for operationalizing the AICAS strategy (Attention, Interest, Chat, Action, Share), adapted in this study from the AISAS model originally proposed by Dentsu [47]. By substituting "Search" with "Chat," the model reflects dialogic, real-time inquiry in CAI–VR environments. These judgments are reinforced by Anaya-Sánchez et al. [51], who confirmed that the clarity of immersive structuring and intensity directly shape user perceptions of destination image and travel intention. Similarly, these technical elements are consistent with Rafi et al. [31], who demonstrated that navigability and UI quality strongly influence behavioral outcomes, and with Calisto and Sarkar [11], who underscored the necessity of embedding interactivity into coherent system architectures.

Turning to RQ4, the findings revealed high stakeholder intention to adopt the model (M = 4.33, SD = 0.65) among both public-sector officials and private-sector practitioners. Regression analysis identified technological self-efficacy, perceived interactivity, and perceived tourism benefits as significant predictors, collectively explaining 73.3% of the variance ($R^2 = .733$). These predictors align with Davis [15] on TAM, Venkatesh et al. [16] on UTAUT, and the self-efficacy construct developed by Compeau and Higgins [19] and later refined by Compeau et al. [21]. They also echo Anaya-Sánchez et al. [51], who showed that perceived immersion strengthens both destination image and behavioral intention, underscoring why interactivity and perceived benefits emerged as significant predictors in this study. Together, these findings confirm that stakeholder acceptance of the 4Ds Model is firmly grounded in both classical and contemporary theories of technology adoption.

Beyond answering the research questions, this study contributes more broadly to academic discourse and system design practice. Theoretically, it integrates VR, CAI, and human-centered design into a unified process model that links requirements analysis, architectural specification, iterative development, and deployment strategy, consistent with system design methodologies [22], [37]. This integration responds to Mariani et al. [34] and Slater and Sanchez-Vives [28], who emphasized bridging immersive depth with structured frameworks. Practically, the 4Ds Model provides policymakers and developers with a technically actionable roadmap to translate contextual needs into architectural components, immersive experiences, and engagement strategies. Its graded immersion levels [28], [51] and conversational engagement design [47], [48] illustrate operational feasibility while preserving adaptability across contexts.

Nevertheless, limitations remain. The stakeholder survey excluded tourists, limiting direct insights into consumer adoption. While the sample was adequate for regression analysis [26], broader sampling across diverse user groups would strengthen generalizability. Future research should conduct pilot implementations across Thailand's four regions to test system performance, scalability, and integration with analytics platforms [29], [31]. Parallel prototyping with tourists as end-users will provide usability evidence, while cross-border application in ASEAN will test adaptability across diverse cultural and policy environments.

VIII. Conclusion

This study proposed the Conversational AI-powered VR Development Model, referred to as the 4Ds Model, as a structured framework for integrating immersive technologies into Thailand's tourism promotion. The model integrates conversational AI, virtual reality, and human-centered design, supported by five enabling capitals: human, cultural, technological, informational, and financial. It is organized into four sequential phases—Discover, Design, Develop, and Deploy—comprising twelve sub-steps. Specifically, the Discover phase mobilizes enabling capitals to ensure contextual readiness; the Design phase structures experiential elements across content, space, and activity; the Develop phase specifies immersion levels and operationalizes interactive features; and the Deploy phase incorporates the AICAS strategy (Attention, Interest, Chat, Action, Share) to sustain engagement through conversational interaction.

The key contribution of this study lies in establishing a unified, culturally grounded framework that bridges conversational AI and virtual reality for digital tourism development in Thailand. The findings confirm that the 4Ds Model is both theoretically robust and practically applicable. Expert evaluations underscored its contextual relevance, while assessments by public-sector tourism officers, destination management professionals, and practitioners in provincial tourism units indicated strong acceptance. Regression analysis revealed that technological self-efficacy, perceived interactivity, and perceived tourism benefits significantly predict stakeholders' behavioral intention to adopt the model.

Theoretically, this research advances digital tourism and AI–VR integration literature by articulating how immersive communication and user-centered design can co-exist in culturally sensitive contexts. Practically, it provides policymakers and practitioners with a replicable roadmap for designing, implementing, and sustaining immersive tourism initiatives that reflect local identity and community values.

Nevertheless, the study acknowledges certain limitations, including the representativeness of the expert sample, the limited scope of stakeholder testing, and the absence of longitudinal validation in real-world deployments. Future research should pilot and refine the 4Ds Model in live tourism projects, explore cross-cultural applicability within ASEAN, and integrate emerging AI capabilities such as generative agents and multimodal interfaces. Thus, the 4Ds Model serves not only as a validated framework for Thailand's digital tourism transformation but also as a transferable reference for immersive tourism innovation across the ASEAN region.

Acknowledgment

This research was supported by the National Science, Research and Innovation Fund, Thailand Science Research and Innovation (TSRI), through Rajamangala University of

Technology Thanyaburi (FRB67E0913) (Grant No. FRB670027/0168). The authors also gratefully acknowledge Mahidol University for their valuable support in providing personnel and access to critical research databases.

REFERENCES

[1] I. Oncioiu and I. Priescu, "The use of virtual reality in tourism destinations as a tool to develop tourist behaviour perspective," *Sustainability*, vol. 14, no. 7, Art. 4191, 2022.

[2] N. Sousa, E. Alén, N. Losada, and M. Melo, "Virtual reality in wine tourism: Immersive experiences for promoting travel destinations," *Journal of Vacation Marketing*, 2024.

[3] R. Adachi, E. M. Cramer, and H. Song, "Using virtual reality for tourism marketing: A mediating role of self-presence," *The Social Science Journal*, vol. 59, no. 4, pp. 657–670, 2022.

[4] M. Melo, G. Gonçalves, F. Jorge, N. Losada, L. Barbosa, M. S. Teixeira, and M. Bessa, "The impact of virtual reality and biological sex on the promotion of tourist destinations: Effects on destination image, place attachment, and behavioural intention," *Journal of Hospitality & Tourism Technology*, vol. 15, no. 1, pp. 18–36, 2024.

[5] S. Wu and S. Wang, "Exploring the impact of AI-enhanced virtual tourism on tourists' pro-environmental behaviour," *Acta Psychologica*, vol. 253, Art. 104773, 2025.

[6] F. Rejón-Guardia, M. A. García-Sastre, F. Orfila-Sintes, and J. B. Garau-Vadell, "Virtual reality in tourism: Centennials' acceptance," *Tourism Analysis*, vol. 25, nos. 2–3, pp. 335–344, 2020.

[7] H. H. Chang, "Virtual reality, YouTube, or social media? Assessing promotional effects on tourism destination," *Journal of Vacation Marketing*, vol. 28, no. 2, pp. 211–227, 2022.

[8] A. Pestek and M. Sarvan, "Virtual reality and modern tourism," *Journal of Tourism Futures*, vol. 7, no. 2, pp. 245–250, 2020.

[9] S. A. G. Yousefdeh and S. S. Oyelere, "Investigating co-presence and collaboration dynamics in real-time virtual reality user interactions," *Frontiers in Virtual Reality*, vol. 5, Art. 1478481, 2024.

[10] D. A. Guttentag, "Virtual reality: Applications and implications for tourism," *Tourism Management*, vol. 31, no. 5, pp. 637–651, 2010.

[11] M. de L. Calisto and S. Sarkar, "A systematic review of virtual reality in tourism and hospitality: The known and the paths to follow," *International Journal of Hospitality Management*, vol. 116, Art. 103623, 2024.

[12] K. Kieanwatana and R. Vongvit, "Virtual reality in tourism: The impact of virtual experiences and destination image on travel intention," *Results in Engineering*, vol. 24, no. 2, Art. 103650, 2024.

[13] A. E. Sousa, P. Cardoso, and F. Dias, "The use of artificial-intelligence systems in tourism and hospitality: The tourists' perspective," *Administrative Sciences*, vol. 14, no. 8, Art. 165, 2024.

[14] S. Wu and L. Zhang, "An exploration of methods of virtual-reality technology to enhance the tourism experience of non-native English speakers," *Applied Mathematics and Nonlinear Sciences*, 2024.

[15] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989.

[16] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, pp. 425–478, 2003.

[17] J. K. Fatima, P. Ghandforoush, M. Khan, and R. D. Masico, "Role of innovativeness and self-efficacy in tourism m-learning," *Tourism Review*, vol. 72, no. 3, pp. 344–355, 2017.

[18] Y. H. Ho, S. S. Alam, M. Masukujjaman, C. Y. Lin, and S. Susmit, "Intention to adopt AI-powered online service among tourism and hospitality companies," *International Journal of Technology and Human Interaction*, vol. 18, no. 1, pp. 1–19, 2022.

[19] D. R. Compeau and C. A. Higgins, "Computer self-efficacy: Development of a measure and initial test," *MIS Quarterly*, vol. 19, no. 2, pp. 189–211, 1995.

[20] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw, "User acceptance of computer technology: A comparison of two theoretical models," *Management Science*, vol. 35, no. 8, pp. 982–1003, 1989.

[21] D. R. Compeau, J. Correia, and J. Thatcher, "Implications of technological progress for the measurement of technology acceptance variables: The case of self-efficacy," *Information & Management*, vol. 54, no. 8, pp. 101–112, 2017.

[22] R. C. Richey and J. D. Klein, *Design and Development Research: Methods, Strategies, and Issues*, New York, NY: Routledge, 2014.

[23] H. Arksey and L. O'Malley, "Scoping studies: Towards a methodological framework," *Int. J. of Social Research Methodology*, vol. 8, no. 1, pp. 19–32, 2005.

[24] R. A. Krueger, *Focus Groups: A Practical Guide for Applied Research*, 5th ed., Thousand Oaks, CA: Sage, 2014.

[25] M. S. B. Yusoff, "ABC of content validation and content validity index calculation," *Education in Medicine Journal*, vol. 11, no. 2, pp. 49–54, 2019.

[26] S. B. Green, "How many subjects does it take to do a regression analysis," *Multivariate Behavioral Research*, vol. 26, no. 3, pp. 499–510, 1991.

[27] H. Song and S. Lu, "The effect of virtual tourism experience on tourist responses: The lens from cognitive appraisal theory," *Asia Pacific Journal of Tourism Research*, vol. 29, no. 7, pp. 885–899, 2024.

[28] M. Slater and M. V. Sanchez-Vives, "Enhancing our lives with immersive virtual reality," *Frontiers in Robotics and AI*, vol. 3, p. 74, 2016.

[29] S. W. Phoong, S. Y. Phoong, and S. L. Khek, "Virtual reality in the tourism sector: A bibliometric systematic literature review of the development and future trends," *Sage Open*, vol. 14, no. 4, Art. 21582440241290933, 2024.

[30] C. Qu, X. Xia, and B. Lv, "Innovative design of cruise-tourism experience based on VR," *Applied Mathematics and Nonlinear Sciences*, 2024.

[31] K. N. Rafi, M. O. Gani, N. M. H. N. Hashim, M. S. Rahman, and M. Masukujjaman, "The influence of 360-degree VR videos on tourism web usage behaviour: The role of web navigability and visual interface design quality," *Tourism Review*, vol. 80, no. 3, pp. 725–741, 2025.

[32] K. Lertbumroongchai, W. Janratchakool, and V. Nittayathammakul, "Using virtual reality and conversational AI to enhance high school students' intentions to pursue computer science: Insights from the CS RMUTT virtual tour," *Studies in Media and Communication*, vol. 13, no. 3, pp. 87–102, 2025.

[33] A. Wiangkham, K. Kieanwatana, and R. Vongvit, "Journey into virtual reality: Identifying behavioural intentions to use virtual reality in tourism through spectral clustering," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 11, no. 1, Art. 100442, 2025.

[34] M. M. Mariani, N. Hashemi, and J. Wirtz, "Artificial intelligence empowered conversational agents: A systematic literature review and research agenda," *Journal of Business Research*, vol. 161, Art. 113838, 2023.

[35] J. Sultana and R. Iqbal, "A framework for context recognition from voice commands and conversations with smart assistants," in *Proc. IEEE 21st Consumer Communications & Networking Conf. (CCNC)*, 2024, pp. 218–221.

[36] S. Javanmardi, E. Javanmardi, and A. Bucci, "Quantifying drivers of virtual reality acceptance in tourism planning using a grey system theory-based approach," *Journal of Hospitality & Tourism Insights*, vol. 8, no. 4, pp. 1308–1327, 2025.

[37] F. Poux, Q. Valembois, C. Mattes, L. Kobbelt, and R. Billen, "Initial user-centred design of a VR heritage system," *Remote Sensing*, vol. 12, no. 16, Art. 2583, 2020.

[38] U. K. Lee, "Tourism using virtual reality: Media richness and IS successes," *Sustainability*, vol. 14, no. 7, Art. 3975, 2022.

[39] M. Zaifri, H. Khalloufi, F. Z. Kaghat, A. Azough, and K. A. Zidani, "Bibliometric and SLR of AR in tourism (2002–2022)," *Multimodal Technologies and Interaction*, vol. 7, no. 7, Art. 64, 2023.

[40] M. Patel, M. Bhatt, and N. S. Rathore, "Promoting smart tourism by implementing VR and AR," *Asian Journal of Engineering and Applied Technology*, vol. 12, no. 1, pp. 28–34, 2023.

[41] A. Nijholt and D. Heylen, "Multimodal communication in inhabited virtual environments," *International Journal of Speech Technology*, vol. 5, no. 4, pp. 343–354, 2002.

[42] J. Baker, K. Nam, and C. S. Dutt, "A UX perspective on heritage tourism in the metaverse," *Information Technology & Tourism*, vol. 25, no. 3, pp. 265–306, 2023.

[43] A. Marasco and B. Balbi, "Designing accessible experiences for heritage visitors through VR," *E-Review of Tourism Research*, vol. 17, no. 3, 2019.

[44] A. Ali and A. J. Frew, "Technology innovation and applications in sustainable destination development," *Information Technology & Tourism*, vol. 14, no. 4, pp. 265–290, 2014.

[45] J. Porritt, *Capitalism as if the World Matters*. Abingdon, UK: Routledge, 2012.

[46] R. Sharpley, "Tourism, sustainable development and the theoretical divide: 20 years on," *Journal of Sustainable Tourism*, vol. 28, no. 11, pp. 1932–1946, 2020.

[47] Dentsu Innovation Institute, *Dentsu Marketing Insight*. Tokyo, Japan: Dentsu, 2005.

[48] K. Sugiyama and T. Andree, *The Dentsu Way*. New York, NY: McGraw-Hill, 2011.

[49] B. G. Tabachnick and L. S. Fidell, *Using Multivariate Statistics*, 7th ed. Boston, MA: Pearson, 2019.

[50] J. F. Hair, W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*, 8th ed. Boston, MA: Cengage, 2019.

[51] R. Anaya-Sánchez, F. Rejón-Guardia, and S. Molinillo, "Impact of virtual reality experiences on destination image and visit intentions: the moderating effects of immersion, destination familiarity and sickness," *International Journal of Contemporary Hospitality Management*, vol. 36, no. 11, pp. 3607–3627, 2024.

[52] Pathrapoowanun, N., Kwangmuang, P., Siritaratiwat, A., & Lan, N. T. (2025). Enhancing Creative Learning in Thai Classical Dance Education: An Integration of Digital Preservation Framework with Ubiquitous Learning Environment. *Digital Applications in Archaeology and Cultural Heritage*, e00460.

APPENDIX: QUESTIONNAIRE REGARDING FACTORS INFLUENCING THE INTENTION TO USE THE CONVERSATIONAL AI-POWERED VR DEVELOPMENT MODEL FOR TOURISM PROMOTION IN THAILAND

Instructions: Before completing this questionnaire, respondents were required to watch a short video presentation introducing the Conversational AI-Powered VR Development Model for Tourism Promotion in Thailand. The video outlined the model's purpose, structure, and potential applications to ensure that participants clearly understood the context before responding. Responses were measured on a 5-point Likert scale: 1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree.

Variable	Item	Scale
Part 1: General Information		
Demographics	Gender, Age, Region, Education, Occupation, Experience with VR, Experience with AI	Categorical
Part 2: Independent Variable (IV)		
Technological Self-Efficacy (X1)	1. I am confident that I can learn to use new technologies on my own. 2. I can troubleshoot basic technical problems when using devices or software. 3. I can apply technology to support my work or activities. 4. I believe that I have sufficient skills to maximize the benefits of technology. 5. I can adapt quickly to technological changes.	1–5
Attitude towards Technology in Tourism (X2)	1. I believe technology makes tourism more attractive. 2. I feel that technology use in tourism should be encouraged. 3. I have a positive view of integrating technology into tourism businesses. 4. I think technology enhances tourism experiences in Thailand. 5. I support the use of technology in tourism promotion.	1–5
Facilitating Conditions (X3)	1. My organization has the necessary devices (e.g., computers, VR headsets). 2. My organization provides sufficient internet connectivity for smooth system use. 3. I receive technical support when needed. 4. My organization offers training or manuals to learn how to use technology. 5. My organization encourages the use of technology in tourism work.	1–5
Perceived Accessibility (X4)	1. The Conversational AI-powered VR system can be accessed anytime and anywhere for tourism purposes. 2. The system is easy to use for tourism personnel. 3. I can use the system even with limited technical knowledge. 4. The system can be accessed through multiple devices for tourism experiences. 5. The system requires minimal time to start using in tourism work.	1–5
Perceived Interactivity (X5)	1. The system interacts appropriately when presenting tourism information. 2. The system responds immediately to my actions. 3. Using the system feels like two-way communication between providers and tourists. 4. The system can adjust content based on tourist behaviors or interests. 5. The system makes tourism experiences more realistic and engaging.	1–5
Perceived Tourism Benefits (X6)	1. The system increases the attractiveness and value of destinations. 2. The system helps tourists decide on destinations more easily. 3. The system allows providers to deliver more in-depth information to customers. 4. The system motivates actual visits and service use. 5. The system supports tourism revenue and sustainable community development.	1–5
Part 3: Dependent Variable (DV)		
Behavioral Intention to Use (Y)	1. I am interested in learning more to develop skills for using the system effectively. 2. I intend to use the system to support my tourism-related work in the future. 3. I plan to continuously use the system in my organization if given the opportunity and resources. 4. I am committed to applying the system to enhance tourist experiences. 5. I would recommend the system to partners or networks in the tourism industry.	1–5

Roadmap for Emerging Cyberbullying Mitigation: Integrating AI-Based Solutions, Ethics, and Policy

Atif Mahmood[1], Shaik Shabana Anjum[2]*, Umm E Mariya Shah[3]*, Pavani Cherukuru[4]*, Javid Iqbal[5], Sarah Bukhari[6]
Faculty of Data Science and Information Technology, INTI International University, 71800 Nilai, Negeri Sembilan, Malaysia[1]
School of Computer Science-Faculty of Innovation and Technology, Taylor's University, Kuala Lumpur, Malaysia[2]
Department of Finance-Kulliyyah of Economics and Management Sciences,
International Islamic University, 53100 Kuala Lumpur, Malaysia[3]
Department of Information Science, Dayananda Sagar Academy of Technology and Management, Bangalore, Karnataka, India[4]
Department of Data Science and Artificial Intelligence-Faculty of Engineering and Technology (FET),
Sunway University, Kuala Lumpur, Malaysia[5]
Department of Information and Communication Technology, Bahauddin Zakariya University Multan, Pakistan[6]

***Abstract*—Cyberbullying is one of these challenges that are most found among the younger users of social media which affects the mental health. Artificial Intelligence (AI) is rapidly developing and has enormous potential to mitigate cyberbullying. Therefore, this chapter will talk about the role AI has started playing in strengthening the efforts to combat cyberbullying. Cyberbullying includes all forms of deliberate aggressive behaviour that aims to inflict social, psychological or physical pain in a digital space and AI detection technologies have a lot of potential to detect, predict and prevent cyberbullying in real time. Other critical components of the chapter are how the advances in Natural Language Processing (NLP) technologies, machine learning, images and videos, behavioural analytics make AI an emerging innovation to prevent cyberbullying and provide better services in a timely manner. There are positive trends that make it clear how Safer AI can help in improving the safety of future digital environments. More advanced NLP models will be able to identify the nuances of cyberbullying involving indirect attacks and sarcasm. The chapter will also discuss the hazards associated with AI-based solutions, such as privacy, the zero-sum game of AI morality against AI effectiveness, and the importance of explaining and assigning responsibility for every AI decision. It shows how AI is changing our approach to online safety and helps us identify cyberbullying in a variety of media, including text, video and images. This article gives an overview of roadmap for cyberbullying mitigation with the assistance of AI and ethical practices.**

Keywords*—*Cyberbullying; human computer interaction; artificial intelligence; natural language processing; machine learning; content moderation; predictive analytics; online safety; youth protection; mental health; mental illness; cybercrime

I. Understanding Cyberbullying

A. Definition and Categories of Cyberbullying

Cyberbullying is the intentional use of internet communication to harass, menace or humiliate another person. Unlike conventional bullying, cyberbullying originates in the digital space and is thereby widespread and persistent since it can cross boundaries of geography and time [1]. Anonymity of online sites can sometimes give abusers more confidence. Lack of face-to-face contact makes the problem worse because the offenders may not consider the emotional impact of their actions on the victim.

Several studies have looked into the issue in depth; others have broken down the broad term "cyberbullying" into specific types including flaming, harassment, impersonation, cyberstalking, flooding, masquerade, denigration, outing and stalking [2,3].

1) Sending offensive or threatening messages repeatedly is harassment. For example, sending hate emails or messages every day.

2) Flaming means participating in rude internet arguments. For instance, commenting offensive things in comment sections.

3) Impersonation is when you assume the identity of another person to harm their reputation. For example, setting up a fake social media account and posting inappropriate content under someone else's name.

4) Cyberstalking is when you continuously monitor or follow someone online with malicious intent to intimidate or control them from their social media activity.

5) Flooding in digital worlds is when someone sends too many unwanted messages, emails or comments to someone. Intended to scare, annoy or silence the target, this is another form of harassment.

6) Masquerade is when someone pretends to be another online to damage relationships or reputation. This can be creating fake profiles, forwarding damaging messages pretending to be someone else, or acting dishonestly to undermine the trust or reputation of the victim.

7) Denigration is the spreading of false or defamatory online information to discredit someone. This can be abusive comments, lies or gossip sent on posts, comments or messages.

8) Outing is when you publicly share private, sensitive or personal information without consent. This can affect the victim's personal or professional life or shame and distress.

9) Cyberstalking is continuous, invasive monitoring or harassment of a person online. It usually means monitoring the victim's activity, sending continuous negative messages or instilling fear through digital means.

*Corresponding authors.

B. Social, Psychological and Physical Effects of Cyberbullying in Digital Spaces

Cyberbullying has a huge impact on people's social, psychological and physical lives [4]. Often those who are socially isolated retreat from both online and offline events because of fear of being ridiculed or more harassment. This disengagement can ruin personal and professional reputation, and affect jobs and social networks and other areas. You can find many more negative psychological effects. Many times, victims suffer from low self-esteem, anxiety and depression. Studies show that cyberbullying victims show symptoms of depression twice as much as non-victims [5]. Persistent long-term stress from cyberbullying can lead to mental health problems like headaches, fatigue, sleep disturbances, high blood pressure and heart conditions. Extreme cyberbullying can increase the risk of suicidal thoughts and self-harm especially in younger population. In extreme cases cyberbullying has led to suicides [6].

These social, psychological and physical effects are more harmful to vulnerable groups, that's why we need to address and prevent cyberbullying to protect ourselves.

C. Impact of Cyberbullying on Younger Social Media Users

Cyberbullying is particularly common among younger social media users, such as teens and preteens, due to their developmental stage and reliance on social recognition [7]. Negative online interaction can have a significant impact on such individuals. Bullied children may skip school to avoid encountering their oppressors, which could result in a decline in academic performance and an increase in absenteeism [8]. Negative online behaviour including creating fake profiles to avoid being found out could be due to fear of rejection or exclusion. Teenagers' growing brains make them more susceptible to mental illness from internet addiction [9]. Public humiliation on sites like Instagram can totally crush people's self-esteem. Studies show that because of its focus on visual content and social validation, sites like Instagram and TikTok are hotspots for cyberbullying [10,11]. A 2021 survey found that almost 60% of kids said they experienced some form of online harassment [12].

D. Current Global Efforts to Combat Cyberbullying

Among the many approaches to cyberbullying are legislative actions, technological innovations, educational programs and support systems. These programs help victims and reduce online abuse.

Legislative actions are needed to combat cyberbullying globally and nationally. The US, UK and Australia have already passed laws on this. The Australian "e-Safety Commissioner" can help reduce online abuse by removing harmful internet content [13]. Meanwhile the European Union's Digital Services Act has a comprehensive plan to increase member internet security. This proposal requires online platforms to limit harmful content so they are accountable in keeping the digital space healthy [14,15].

Partially technology has helped reduce cyberbullying. On social media networks like Facebook, content moderation tools and reporting mechanisms allow users to flag objectionable content. Facebook uses artificial intelligence to aggressively search for hate speech and inflammatory language [16]. Meanwhile artificial intelligence-powered platforms like Rethink give consumers quick cues to rethink before they share potentially dangerous content [17].

The fight against cyberbullying is mostly focused on educational activities. Non-governmental organisations (NGOs) and educational institutions are implementing projects to increase digital literacy and empathy among young people. The "Delete Cyberbullying" project teaches teens about the effects of online harassment (https://www.endcyberbullying.net/). Meanwhile events like Safer Internet Day brings together educators, politicians and tech companies to discuss practical solutions and raise awareness on cyberbullying.

Support systems are needed for cyberbullying victims to manage their psychological and emotional response to online harassment. Helplines and initiatives like Childline (https://childline.org.uk/) in the UK are available for victims. Meanwhile social media platforms have introduced tools like Instagram's "Restrict" feature which allows users to block interactions with potential attackers so victims have some sense of security [16]. These combined strategies are a total solution to cyberbullying.

E. Limitations of Current Approaches

Despite all the work, current approaches to cyberbullying have big flaws that make them not work. Legislative is a main challenge. Cyberbullying is global, law enforcement projects are across borders. Tracking down anonymous offenders across different countries is a real challenge for law enforcement.

To make it worse, many social media platforms rely on users to report cyberbullying. This means delays in response to incidents [18]. Many times, victims complain about platforms not responding to objectionable content. Many times, algorithms used to detect negative behaviour can't understand context, sarcasm or cultural sensitivity. AI systems can flag innocent banter as harassment so making it harder to detect cyberbullying.

The shame of cyberbullying stops victims from seeking help. Many delay reporting because of fear of being labelled as "oversensitive" or retaliation. This stigma stops victims from getting the help and support they need [13]. Plus, there are no educational initiatives to increase awareness and stop cyberbullying. Many projects don't keep up with the fast pace of digital interactions so new online environments like BeReal don't have enough security and user services. These limitations highlight the need for continuous work to reduce cyberbullying.

F. The Importance/Potential of AI in Mitigating Cyberbullying

As cyberbullying has emerged as a significantly widespread problem, it is required to adopt the efficient detection and prevention measures. Artificial intelligence is essential in tackling cyberbullying. It provides innovative and adaptable solutions that are scalable as well. It facilitates in real-time monitoring and reviewing the content for automated detection and classification of hazardous messages. Thus, assisting in the identification and reduction of cyberbullying events.

Furthermore, AI can also help to increase awareness among digital space to combat with cyberbullying.

This paper discusses the importance of AI in cyberbullying mitigation in Section I followed AI-powered solutions for prevention of cyberbullying in Sections II and III respectively. The future recommendations and directives have been elaborately discussed in Section IV followed by concluding remarks in Section V.

II. AI-Powered Solutions for Cyberbullying Prevention

AI can be utilized to identify cyberbullying in social networks. It can offer a range of solutions to combat cyberbullying. Such solutions are comprised of real-time detection and intervention, predictive modelling and risk assessment, content moderation and filtering and victim support. Moreover, AI-based educational tools and personalized learning approaches could be incorporated to spread awareness and educate individuals to address such situations and make decisions [19].

AI solutions are made up of text analysis, multimedia content analysis and user behaviour analysis as depicted in Fig. 1. Natural language processing (NLP), machine learning and data mining are used in AI powered cyberbullying solutions. These can help build a robust cyberbullying detection system.

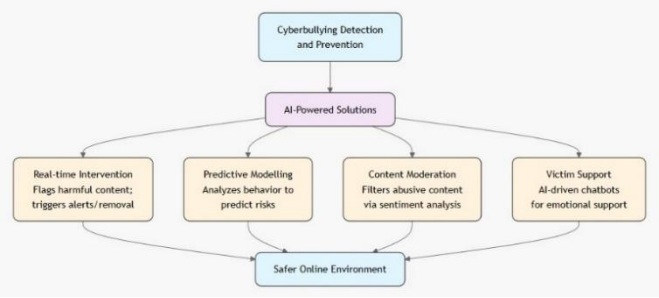


Fig. 1. AI powered solution.

A. Real-Time Intervention Mechanisms

For text analysis, linguistic insights can be extracted using NLP. It will help identify taunts, insults, threats and other harsh or abusive language in the text.

Multimedia content is audio, video and image content so cyberbullying can be detected through various methods. Images and videos can be analysed by AI algorithms involving facial expression recognition, body language analysis, object recognition and image manipulation detection [20]. Audios can be analysed for tone, pitch and volume variation and recognising sound relevant to violence or distress. Combining text, audio and visual indicators in multimodal analysis helps to understand the background and intention of the content [16].

Datamining can help to reveal the broader context of human behaviour to uncover hidden patterns related to cyberbullying. It involves examining user profiles, posting behaviour of the user, interaction pattern and social contacts to identify possible victims or cyberbullies [21].

In this way text messages, online comments and social media posts could be automatically flagged by the AI-powered system. Moreover, several interventions could be triggered to combat cyberbullying like sending alerts to the human reviewers or moderators, warning notifications to the offenders, automatic removal of the hazardous content, connecting the victims to the support channels, helplines or counselling services [22].

B. Predictive Modelling and Risk Assessment

Proactive prevention of cyberbullying relies much on the AI-driven predictive modelling and risk assessment strategies. These strategies observe the online activities and interaction pattern to analyse user behaviour. Based on the analysis, system can predict who would be more likely to engage in cyberbullying as an offender or a victim. Such proactive approach enables platforms to take necessary measures to minimise the damage thus fostering a safer online community.

Machine learning could help to develop predictive models based on predefined labels and large datasets that can classify and detect cyberbullying based on word frequency, the use of particular emojis, or the overall message sentiment [23]. Nevertheless, the scarcity of publicly available dataset, especially having multimedia content and the limitation of several studies based on only textual data poses a significant challenge in the development of cyberbullying detection systems capable of analysing images and videos [20]. Furthermore, there is no standardized definition of cyberbullying and are no specified rules or guidelines for data labelling as well. This could also lead to inefficient cyberbullying detection algorithm.

III. Advances in AI Technologies for Cyberbullying Prevention

Artificial intelligence has emerged as a powerful tool in combating the growing problem of cyberbullying [24]. Recent developments in deep learning and natural language processing have enabled the creation of sophisticated models that can effectively detect and mitigate the impact of cyberbullying incidents [25]. One key area of focus is the use of sentiment analysis and behavioural analytics to identify potential cyberbullying threats. By analysing the tone, context, and language patterns of online interactions, AI-powered systems can flag concerning content and alert authorities or community moderators, allowing for timely intervention.

Moreover, machine learning algorithms have garnered attention for their ability to moderate user-generated content, filtering out harmful or abusive material and curbing the spread of cyber-violence. Beyond detection, AI techniques are being leveraged to enhance the proactive prevention of cyberbullying. Intelligent agents and chatbots, powered by natural language generation, can engage with users at risk, offering emotional support, guidance, and strategies to build resilience.

A. Natural Language Processing (NLP)

1) Text analysis and sentiment analysis: Utilizing NLP techniques to identify bullying language patterns, tone, and context. Recognition and classification of dangerous or abusive content on websites and social media [26]. Creating sophisticated deep learning models to decipher the subtleties of

language used in online conversations, such as transformer models and recurrent neural networks [26]. Although natural language processing has been a major area of interest, scientists are investigating a variety of cutting-edge strategies in the quickly growing field of artificial intelligence for cyberbullying prevention [27].

2) Identifying cyberbullying patterns and language: Examining textual data to find signs of cyberbullying, including aggressive exchanges, derogatory language, and emotional trends [24,26]. Recent developments in deep learning architectures, like transformer models and recurrent neural networks, have shown impressive potential in precisely identifying and categorising cyberbullying content. Researchers have been able to build systems that can detect and distinguish cyberbullying from regular online interactions by using these models with big data [28, 29,56].

3) Recognizing sarcasm and indirect attacks: One of the challenges in detecting cyberbullying is finding indirect or sarcastic forms of harassment which can be more subtle and harder to identify. To detect these more advanced forms of cyber aggression AI powered systems are being developed to detect contextual information and subtle language cues.

To better understand the cyberbullying landscape some studies have looked into multifaceted approaches, combining text analysis with other sources of information like images and videos [24].

B. Machine Learning

Machine learning is being used more and more in risk identification and content moderation to prevent cyberbullying. By looking at user generated content these algorithms can identify and remove offensive or dangerous content and stop cyberviolence from spreading.

By looking at online trends and interactions behavioural analytics has also been explored as a way to predict who will experience cyberviolence. Scientists have looked at various feature sets to develop strong machine learning models for early detection of cyberbullying incidents - content driven, profile based, mobile app data etc [30].

More precise and effective cyberbullying detection systems have been made possible by machine learning. To use the power of neural networks in text data analysis and cyberbullying pattern recognition researchers have tried out various deep learning algorithms - Gated Recurrent Units, Long Short-Term Memory, Bidirectional LSTM.

1) Predictive modelling to prevent cyberbullying: Machine learning-based predictive models were also implemented to the cyberbullying prevention system. These models can help predict people at risk of being-involved in CB from past patterns and information.

Based on this group of at-risk individuals, targeted interventions, such as AI-based interventions, including personalised support, counselling, and education materials, can be provided to help them build resilience and encourage positive online behaviours [25, 30].

One promising approach to address the challenging problem of cyberbully prevention is integrating the proactive intervention techniques and AI-based predictive modelling [29].

2) Data analysis-based methods for user behaviour study: In order to obtain better cyberbullying detection, researchers have been studying data-driven approaches that exploit peer influence and per user feature set beyond the well-known detection and classification strategies. In order to offer a more accurate and personalised detection of cyberbullying, these tailored models can also take into account a user's individual online activity, social media profile and interaction specificity [31].

These AI-based systems can recognize new trends in cyber-heists and also act more quickly by incorporating more granular data about users as well as peer-to-peer dynamics. This could halt the dynamics of the situation from exacerbating and reduce harm to victims [30].

C. Image and Video Analysis

While most AI-based cyberbullying prevention related work has focused on text analysis, it is increasingly recognized that these approaches need to be extended to other modalities, like images and videos. The emerging area of computer vision and multi\-media analysis have been employed by researchers to address visual facets of cyberbullying such as circulation of embarrassing or humiliating images and videos. Deep-learning algorithms also trained on massive numbers of cyberbullying-related photos and videos can help locate and flag harmful media, leading to pre-emptive measures and better content moderation.

AI-powered systems that integrate sophisticated text analysis and visual recognition capabilities provide a more complete method for tackling the complexity of cyberbullying by covering both written and visual aspects [32].

1) Detecting harmful content in multimedia: The increased presence of user-generated multimedia content, such as videos and images, has also provided a new challenge in detecting instances of cyberbullying. To address this issue, researchers have several projects that develop AI solutions to analyse visual data to identify or detect potentially harmful or abusive content. These systems can use various computer vision approaches including object detection, face recognition or detection, and scene understanding to detect and identify images or videos that could potentially be representative of a cyberbullying related incident.

2) Detecting cyberbullying through visual cues (e.g., body language, facial expressions): In addition to identifying harmful media, there are a few AI-powered systems that are currently investigating using visual cues to detect cyberbullying behaviour. By analysing the context, body language, and emotional expressions in images and video, these AI models can be leveraged to identify patterns of behaviour that can act as indicators of cyber-aggression, even when there is no text content [27][57].

The combination of visual-based analysis with a traditional text-based approach may result in a more comprehensive picture of the cyberbullying phenomenon, which can help inform more effective detection and intervention protocols [33].

D. Behavioural Analytics

Behavioural analytics have taken root as an instrumental resource in the prevention of cyberbullying by utilizing AI to assess patterns of online interactions and behaviour. Algorithms focused on behavioural analytics consider the frequency of communication, tone of communication, and social networks, which are valuable predictors of whether individuals are defenceless or if their behaviour is a risk for cyberbullying. The combination of AI-driven behavioural analytics and personalized intervention approaches has potential to empower victims, support positive online communities and encourage citizenship online.

1) Analysing user behaviour patterns to find at-risk cyberbullying victims or perpetrators: Research has been done which aims to harness AI-powered behavioural analytics to discover early warning signs of cyberbullying such as changes in online activities, communications, and social network interactions. AI can help monitor and analyse user behaviour data to discover individuals who may be prone to cyberbullying victimhood or may display concerning behaviours which could escalate into cyber-aggression.

In even a preliminary use of behavioural analytics to detect potential for harm, early identification of these behaviour patterns could facilitate timely interventions with at-risk individuals leading to connection with appropriate supports and resources to help them build resilience and positive online engagement.

2) Tracking digital interactions to identify aggression and conflict: More than just identifying victims and offenders - AI behavioural analytics can be used to track interactions in digital spaces for signs of aggression, conflict, and rising tensions. If AI systems analyse tone, sentiment and dynamics of the digital conversation, they may show early warning signals for cyberbullying so the inventible mechanism can kick in. This proactive data-driven model to moderation and community management can completely alter how we consider - and respond to - cyberbullying and increase levels of acceptability of online spaces.

To summarize, where there is great potential for preventing cyberbullying is through radical data taking using AI-powered solutions. More specifically, there are an unlimited number of opportunities for real-time monitoring of online environments whether through imaginative text analysis methods, multimedia recognitions or through developing personalized behavioural analytics and access to the patterns of users across platforms, AI holds a lot of promise to help detect, prevent, and intervene, when being victimized online [24,25,34].

Practitioners and scientists may want to consider how they can leverage AI in developing digital wellbeing interventions that may result in much more efficient, scalable, and adaptable services. They can also bring some responsibility back to owners of digital spaces to create more inclusive and safer online communities and places where it is safe to innovate [24, 34, 35, 36].

E. Challenges and Considerations

There is tremendous promise in AI-Supported cyberbullying prevention, although researchers and practitioners have found several important challenges and considerations. Among the most important are ethical issues, privacy implications, and needing different disciplines to work together to substantiate responsible and effective uses of these technologies.

F. Privacy Considerations

The collection and analysis of user data for the detection of cyberbullying and for intervention raise serious privacy implications. AI systems need to have strong privacy and security protocols in place to protect personal information, particularly for minors who could be vulnerable to cyberbullying [37].

In order to establish user trust, and ensure compliance with privacy laws, it is important to deliberately consider the processes for collecting, storing, processing, and being transparent and accountable about how data is collected and used by these systems [24].

1) Data collection and use for AI models: When developing AI-based cyberbullying prevention and detection tools, it is often important to collect and analyse user data, such as information from online chats, online social networking communications, and patterns of digital behaviour [38].

This information has to be processed with utmost care and vigilance to protect personal privacy and avoid possible abuse or exploitation—even if it is critical for training and validating these AI systems. In order to ensure that personal information is only used for lawful and ethical purposes, developers and researchers must implement strict data management frameworks that provide clear rules on data collection, user consent, and data de-identification [39].

2) Safeguarding user privacy and online anonymity: Consideration for personal safety and online anonymity should be explored when developing AI systems to prevent cyberbullying, especially in regard to children and youth (i.e., vulnerable users). It is important chatbot and AI systems do not breach user privacy and jeopardize the anonymity many users need to feel safe, secure and empowered to navigate their online lives [34].

Developers must provide a fair balance in protecting user privacy and online safety while identifying and addressing cyberbullying incidents.

G. Ethical Considerations

The use of AI-powered solutions to combat cyberbullying raise a variety of ethics-related questions that will need to be taken seriously. One question to consider is the obligation to ensure that these advancements are developed and deployed in ways that honour the fundamental rights of individuals and cultural norms, along with requisite algorithmic calculation, and the concern of unintended consequences.

1) Algorithmic bias and fairness: Cyberbullying prevention measures must be thoroughly assessed for bias and fairness to ensure that they do not unjustly single out or disfavour specific people or groups. Discriminatory judgements that further disenfranchise already marginalised groups and perpetuate complaints can be caused by biases in algorithmic decision-making, model design, or training datasets [40].

2) Transparency and accountability in AI-enabled decision-making: AI-enabled cyberbullying control or prevention systems must be transparent and accountable to users, stakeholders, and regulators for its decisions and actions. Processes and procedures must be put in place, and detailed, to evidence, show, and or demonstrate the workings of the system, the decision processes, and how the system responds to instances of cyberbullying.

In the absence of similar transparency and accountability, these technologies may be perceived as opaque, unreliable, and perhaps subject to abuse or inappropriate misuse [41].

3) The potential for AI to be repurposed for malicious ends: The aim of the AI-powered cyberbullying prevention programs is to render online spaces safer and comparatively friendly, but there is a possibility for them to be misappropriated for illegal ends such as censorship, surveillance, or to purposely harass certain individuals or groups. To ensure that these technologies are utilized for appropriate and ethical purposes and do not infringe upon fundamental human rights or democratic freedoms, protections and governance structures must exist [42].

H. The "Zero-Sum Game" of AI Morality and Effectiveness

There might be a disjunction between the moral and ethical consequences of AI's expanding abilities to detect and respond to cyberbullying and the efficiency of systems employed to address this issue. In employing AI-enabled cyberbullying surveillance systems, there must be a cognitive divorce between AI tools preventing cyberbullying with moral and ethical soundness with high expectations of privacy, equity, and protection of human rights on one side and the use of AI tools for efficacy and convenience against cyberbullying on the other side.

However, the more successful these systems are at detecting and addressing instances of cyberbullying, the greater amount of user information may be required to gather and examine, which could cause privacy issues. This "zero-sum game" between AI efficacy and morality is a serious issue that needs to be resolved by careful, multi-stakeholder cooperation and the creation of strong governance structures.

AI-powered preventative cyberbullying methods that are socially conscious and have an impact on safer online environments will only be successful if the proper balance is struck between ethical considerations and technological efficacy [24, 25, 43, 44].

In the end, preventing cyberbullying with AI will need a multipronged strategy that emphasises both modern technology and moral user-centred design. Researchers and developers can endeavour to create solutions based on AI that are not only successful at recognising and reducing cyberbullying but also in line with fundamental liberties and societal values by tackling the important issues concerning information privacy, bias within algorithms, and transparent accountability [24, 44, 45, 46].

1) Balancing the need for accurate detection with the risk of false positives: There is always a chance of false positives, in which innocuous or normal interactions are mistakenly reported as cyberbullying, even though AI-powered cyberbullying detection systems can be very successful at spotting potentially harmful content or behaviours.

The AI-based cyberbullying solutions could additionally restrict applicable people or groups or unintentionally over-censor acceptable behaviours suppressing acceptable amounts of online expression.

In order to more effectively identify cyberbullying while avoiding false positives that may limit rights and liberties of users, developers must maximize the sensitivity and specificity of their AI outcomes.

To find the right balance and ensure that AI-based cyberbullying prevention solutions are effective and socially responsible will require, significant and extensive user testing, iterative model improvement, and close collaboration with a broad range of stakeholder groups [47].

2) Striking a balance between user privacy and the effectiveness of AI: While a simple conceptual model of a cyberbullying detection and intervention system based on AI could be developed, the "effective" part of a system likely requires access to a more extensive data set including social networking activity, online communications and perhaps even personal data.

Even if it is done to increase online safety, users might think that their private data has been inappropriately used or excessively exposed, which is why the gathering and utilization of such sensitive data impacts serious privacy concerns.

As a result, developers must carefully balance the requirement for efficient AI-powered cyberbullying detection with users' fundamental right to privacy and control over their personal information. This could entail creating privacy-preserving strategies like federated learning or differential privacy in addition to strong data governance frameworks that give users the ability to control how their data is used. The future of AI-powered cyberbullying prevention may be one of greater efficacy, social responsibility, and user trust if these major issues are resolved [25]. The summary of findings for AI-powered Cyberbullying mitigation has been tabulated in Table I.

TABLE I. SUMMARY OF FINDINGS: AI-POWERED CYBERBULLYING MITIGATION

Aspect	Key Technologies/Methods	Primary Applications & Capabilities	Identified Challenges & Considerations
AI Modalities & Solutions			
Natural Language Processing (NLP)	- Transformer Models (e.g., BERT, GPT) - Recurrent Neural Networks (RNN, LSTM) - Sentiment & Contextual Analysis	- Analysis text to detect abusive language, threats, and insults. - Identifies patterns of harassment and denigration. - Aims to understand sarcasm and indirect attacks.	- Difficulty with linguistic nuance, cultural context, and sarcasm. - High risk of false positives/negatives without deep contextual understanding.
Machine Learning (ML) / Predictive Analytics	- Supervised & Deep Learning (GRU, Bi-LSTM) - Behavioural Pattern Analysis - Peer Influence Modelling (e.g., PI-Bully)	- Classifies cyberbullying based on word frequency, emojis, and sentiment. - Predicts at-risk users (potential victims or perpetrators). - Enables proactive, targeted interventions.	- Scarcity of large, labelled, and multi-modal (text, image, video) datasets. - Lack of a standardized definition of cyberbullying for data labelling.
Image & Video Analysis	- Computer Vision - Facial Expression & Body Language Recognition - Object & Scene Recognition	- Detects harmful visual content (embarrassing images/videos). - Identifies visual cues of aggression or distress. - Multimodal analysis combines visual and textual context.	- A relatively under-explored area compared to text analysis. - Requires complex models to integrate visual and contextual data.
Behavioural Analytics	- User Interaction Analysis - Social Network Analysis	- Tracks digital interactions to identify early signs of aggression and conflict. - Analyses user profiles, posting frequency, and social connections to flag risky behaviour.	- Raises significant privacy concerns due to continuous monitoring. - Requires balancing intrusion with effectiveness.
Core Challenges			
Ethical & Privacy	- Data Governance Frameworks - Transparency Protocols	- Ensures user data is protected, especially for minors. - Aims for accountable and explainable AI decision-making.	- Privacy Risk: Collection of sensitive personal data. - "Zero-Sum Game": Tension between AI efficacy (needing more data) and user privacy/morality (protecting data).
Algorithmic Bias & Fairness	- Bias Mitigation Strategies - Fairness Audits	- Seeks to create equitable systems that do not disproportionately target or harm marginalized groups.	- Biased training data can lead to discriminatory outcomes and reinforce existing societal inequalities.
Effectiveness & Accuracy	- Model Optimization - Iterative Testing	- Aims to maximize accurate detection of true cyberbullying incidents.	- False Positives: Risk of censoring innocent banter or legitimate speech. - False Negatives: Risk of missing sophisticated or novel forms of harassment.
Future Directions			
Technical Innovations	- Advanced NLP for subtlety - Real-time monitoring dashboards (e.g., AI Ally) - Personalized AI (e.g., CAPTAIN chatbot) - Cross-platform tools (e.g., Bullstop)	- Improve detection of nuanced attacks. - Enable immediate intervention. - Provide tailored support to victims. - Create a unified defence across social media, messaging apps, and forums.	- Requires significant R&D investment and interdisciplinary collaboration.
Policy & Collaboration	- Multi-stakeholder Frameworks	- Bridges the gap between researchers (evidence), policymakers (regulation), and industry (implementation).	- Creating effective, enforceable, and adaptable policies that keep pace with technological change.
Education & Awareness	- AI-Powered Chatbots & Serious Games - Data-Driven Educational Content	- Provides interactive, personalized learning about cyberbullying effects and resilience. - Empowers users with knowledge and coping strategies.	- Ensuring widespread adoption and cultural relevance of educational programs.

IV. FUTURE DIRECTIONS AND RECOMMENDATIONS

A. Innovations Required to Make AI More Effective in Cyberbullying Prevention

Considerable advancements in the field of artificial intelligence (AI) have been recorded over the past few years, but none of them are sufficient in preventing cyberbullying. The main weakness is in the area of natural language processing (NLP). Modern AI may notice the boldest attacks, but even it will always fail to realize the milder strategies- particular moments that foster bullying or understated sarcasm that is implemented with the expressive intention to harm. NLP improvement would thus play a critical role towards addressing such blind spots. It is empirically proved that, under such exposure to a large set of examples, the AI models are shown to have an increased ability not only to detect blatantly inappropriate expressions but also to identify the malicious patterns that are repeatedly used by harassers [48].

One of the most telling developments is the ability of AI systems to follow real-time online activity and act in a timely manner. Continuously tracking what happens, computational agents are now able to identify dangerous interactions and act before things go out of control. One such example is AI Ally dashboard that monitors Discord and other platforms. It allows users to identify the wrong exchange and to collect proof of the cyberbullying [49]. This is a tool that will then allow more effective counteractions and this will enhance defences against abuse. In a more practical sense, cyberbullying is alleviated

through AI personalisation: the use of individualized assistance that would be shaped to suit specific victims has proven to be significantly more efficient as compared to uniform, one-size-fits-all interventions. The case of CAPTAIN chatbot demonstrates this concept because the chatbot adjusts discourse on the fly and supports victims who fight against cyberbullying. Since these AI tools are environment-dependent and tuned to the requirements of respective users, such a tool poses a significantly greater resource than generic, standardized alternatives [50].

Colleagues and students, first, it can be noted that there is nothing exceptional with the urgency of developing effective technology to mitigate cyberbullying in a particular social-media ecosystem; instead, what is necessary is that the robust solution be platform-agnostic. Stated differently, the tool has to work harmoniously with a variety of social-networking platforms, instant messaging programs, and Internet message boards. A distinctive suggestion that has been made in this area is the Bullstop application, which will attach to already existing social-media sites and will build a shield against online bullying. Constantly scanning user feeds checking on objectionable content, Bullstop can use this to give warnings and can recommend ways that a bullying event can be defused before it escalates [48].

Ethical imperatives form a preset prerequisite in the academic field of enquiry of creating Artificial Intelligence capable of preventing cyberbullying. The transparent interface, privacy of the user, and maintaining ethical standards require careful implementation into a technological solution. The informed consent, the maintenance of clear and transparent data-management procedures, and the adoption of the mechanisms to help eliminate the misuse or the interruption of the system itself, are essential. AI systems can only be trusted to work effectively against cyberbullying when these ethical dimensions are comfortably tackled.

With the help of prioritizing such forward-thinking of ethical approaches, modern AI technologies will be able to improve their potential to maximize the possibilities of harmful online behaviour identification, alleviation, and elimination. Such development of AI would allow at once to increase the safety and supportiveness of online environments to all users in addition to enhancing the capacity of the system to identify and address potentially dangerous conduct.

B. Collaboration Between Researchers, Policymakers, and Industry Stakeholders

Researchers, stakeholders in the industry, and policymakers need to work together in order to reduce cyberbullying with the help of artificial intelligence in our modern world. Practical approach to the problem of online harassment detection and prevention is entirely in the domain of research organizations. In this case, the academic community also needs to develop a new architecture of algorithmic solutions, conduct empirical studies, and provide evidence-based recommendations on the responsible use of AI-powered technologies [48]. Legislators, however, play a pivotal role in the development of regulation governing the application of AI in the online environment. They would be responsible to oversee the actions of AI systems in terms of moral regulations, user privacy, and taking ownership of the possible negative consequences [49]. There are ethical dilemmas or an abuse issue that can occur without a proper policy framework of using AI assistance from AI-based tools.

The wide acceptance of AI technologies will be based on the stakeholders in the industry, especially those belonging to the tech companies' section, messaging, and social media platforms. They will be in charge of incorporating these technologies in their systems and making sure that they can be used. The efficiency of cyberbullying prevention using AI will rely on the implementation of the results of the studies and the compliance of the stakeholders of the industry to the policies [50]. These stakeholders can assist in achieving a safer online space by making sure AI solutions are practical, scalable to a reasonable degree and meet the legal requirements.

This cooperation results in a strategy to stop cyberbullying made up of the various areas. Policymakers make sure that ethical and legal issues are considered, scientists deliver the groundbreaking information and researches, and industry authorities implement these solutions. In collaboration, they can create the system which will help to popularise AI technologies and make them both ethically right and productive.

C. Education and Awareness Programs for Users and Online Communities

Cyberbullying needs sufficient education and awareness programs to provide users and online communities with awareness of cyberbullying, and related knowledge to approach and respond to it. Besides creating awareness, such programs should be crafted to impart the needed skills in users on how to handle bullying in online environments. Such efforts have a potential to attract a substantial number of people and leave a lasting impact on the reduction of cyberbullying behaviour through the use of state-of-the-art strategies, such as artificial intelligence (AI).

One of the AI tools which is rapidly becoming an effective means of avoiding cyberbullying is the CAPTAIN (Cyberbullying Awareness and Prevention Through Artificial Intelligence) chatbot.

The CAPTAIN chatbot provides users with personalised knowledge on the negative effects of cyberbullying and is available on-demand when the user requires assistance. They can also practise the responses and perfect it as through simulations of actual cyber bullying instances, the users can become more prepared and confident to handle should it arise online [50]. Having conversational AI in the educational program, the users will be offered with a personalised experience in which they will be able to receive preliminary measures that will match their unique needs and, therefore, will shift towards higher effectiveness.

The application of AI-enabled serious games is another good strategy since they present an entertaining approach to learning about the negative effects of cyberbullying. It is possible to achieve high-levels of user engagement (an extended interaction) through the use of serious games because, by incorporating game mechanics to educate people on relevant social topics in an entertaining, interactive form; they advance engagement with greater results. Such games can also employ AI to customise the experience of each player by either varying

the content or changing the level of difficulty depending on the actions of the player. Such interactive learning also allows users to understand the practical effects of cyber bullying as well as promoting empathies and respect in online communication [51]. Moreover, the games give room to the players to practice different scenarios and create an emotional intelligence they would require in events of being cyberbullied.

AI can also have a high level of utility in the creation of specific educational materials through the analysis of user activity and preferences. This will be a data-driven measure, which makes the content that is going to be used valid and suitable to the requirements of people with different users, which allows creating the educational content that will be appealing to this or that demographic group. As an example, a gamified learning content might be of greater value to younger users, and case studies and detailed discussions might be of greater value to senior users [52]. Optimizing the educational contents based on the user data, AI guarantees that awareness campaigns will reach a broad audience, not confined to users of a particular age group and cultural background.

Combined with these AI-based tools, outreach to social networks is a crucial part of achieving education and awareness campaigns. Campaigns can be promoted through the social media platforms, online forums and other virtual platforms, educational materials can be disseminated and positive online behaviour can be empowered. Peer-to-peer support is also another very important element of these networks since users may provide each other with tips and techniques to be successfully introduced to fight cyber bullying. Ensuring the community-based approach to education, awareness campaigns could be more open and effective and allow users to participate in some proactive steps in cyber bullying prevention.

Finally, AI can be implemented in an educational and awareness program, which offers new solutions to cyberbullying. Chatbots and serious games are examples of the AI-guided technologies that provide scalable, personalised and engaging solutions that not only created awareness but also grant users confidence to act. Focusing on the needs of users and tailoring information to fit their needs, AI could contribute to making the online world safer and more respectful so that individuals would recognise the harmful effects of cyberbullying and possess the skills and knowledge needed to combat it.

D. Ethical Frameworks and Guidelines for Responsible AI Deployment

Due to the rise in AI technologies deployment, it is crucial to make sure that these systems are developed and implemented responsibly. Ethical principles play a pivotal role in eliminating the abuse of AI in delicate fields such as the elimination of cyberbullying. Such frameworks could act as a guideline to the legislators as well as the developers in coming up with systems that put emphasis on privacy, justice, accountability and transparency.

AI systems, especially the ones preventing cyberbullying should adhere to the principles of ethics in order to safeguard the users. Fairness is one of the keys of these structures. The AI algorithms should also be open and are not supposed to accidentally affirm built-in prejudices. As demonstrated by study [53], ethical AI laws must be able to tackle the issue of fairness by ensuring that AI systems avoid discrimination against certain people based on race, gender, as well as, the socioeconomic status. Based on their observation, AI models trained with biassed data tend to strengthen such biases, which may harm the under-represented population. According to them, the frameworks of ethical AI must have provisions to reduce biases and have ethically fair returns to all users.

User privacy is another basic value of successful AI implementation. The increased interest in data privacy, including the social media, has evinced an issue of handling personal data with care by AI systems. [54] highlighted the importance of data collection and processing transparency and considered the ethical issues of AI used in moderating social media. They suggest that the AI-based platforms can provide more control over the user data and guarantee privacy without interfering with the system in terms of identifying harmful behaviour, i.e., cyberbullying. It [54] states that their paradigm promotes a compromise between protection of user rights and effectiveness of an AI system.

Moreover, responsibility is crucial even when it comes to the creation and implementation of AI systems. The systems developed by AI frameworks should have systems that demand responsibility of the makers of the systems. Developers should make their systems explicable to both users and external auditors in order to be able to understand the decision-making process, according to a study conducted on ethical AI of online platforms by [55]. They stress that since there is a role of AI system developers in the designing of the technology and its consequences, they should also be blamed.

We also need ethical rules which would make sure that we implement AI system in a manner which cannot be misused especially among the already vulnerable groups. As an example, to fight the issue of cyberbullying, one would have to design AI that would be able to detect the harmful behaviour without violating the freedom of expression of users. Guidelines have to consider the need of ensuring that the freedom of expression of the users is not infringed upon and the need to regulate the contents. This is of particular concern to social media sites which, are habitually accused of censoring content in a manner that is discriminative to certain groups of users.

Such structures should be adaptable and versatile due to the development of AI technologies. In the development of AI, as well as enhancing these ethical frameworks, systematic work by the researchers, legislators, and business executives is necessary as stakeholders of the current development of AI. This will ensure that AI remains a good influence that brings about a fairer and a more just digital society.

V. CONCLUSION

The multifaceted nature of cyberbullying, with its considerable social, psychological, and physical effects, particularly on younger social media users, calls for more effective intervention strategies than what global efforts presently can provide. While these measures are quite essential and provide a good foundation, their inadequacy calls for new solutions. With AI, the possibilities are endless to fight

cyberbullying via AI-powered solutions to directly take action in real-time and forecasting models to determine risk. Breakthrough advances in AI-related technologies, such as Natural Language Processing, Machine Learning, Image & Video Analysis, and Behavioural Analytics are providing mind-blowing tools that can be used to comprehend and address this pervasive problem.

However, if the transformative potential of AI is to be truly unlocked in cyberbullying prevention, key areas will require targeted development and innovation. These areas include fostering strong collaboration between researchers, policymakers, and industry stakeholders to ensure that best possible solutions are designed and implemented ethically. Meanwhile, in parallel with the collaboration, there will need to be extensive education and awareness initiatives targeting users and online communities to empower them with knowledge on the identification and administration of cyberbullying. Finally, establishing ethical frameworks and guidelines for the responsible use of AI must take priority to address challenging questions around privacy, bias, and accountability; ultimately paving the way for a safer and more positive online environment for all.

REFERENCES

[1] R. Slonje and P. K. Smith, "Cyberbullying: Another main type of bullying?" Scandinavian Journal of Psychology, vol. 49, no. 2, pp. 147–154, 2007.

[2] S. Bauman, "Types of cyberbullying," in Cyberbullying: What Counselors Need to Know, 2015, pp. 53–58.

[3] S. Mahbub, E. Pardede, and A. S. M. Kayes, "Detection of Harassment Type of Cyberbullying," Security and Communication Networks, vol. 2021, pp. 1–12, 2021.

[4] F. A. Esquivel, I. L. De La Garza López, and A. D. Benavides, "Emotional impact of bullying and cyber bullying," Revista Caribeña De Ciencias Sociales, vol. 12, no. 1, pp. 367–383, 2023.

[5] G. Niu, J. He, S. Lin, X. Sun, and C. Longobardi, "Cyberbullying victimization and adolescent depression," International Journal of Environmental Research and Public Health, vol. 17, no. 12, p. 4368, 2020.

[6] S. Hinduja and J. W. Patchin, "Bullying, cyberbullying, and suicide," Archives of Suicide Research, vol. 14, no. 3, pp. 206–221, 2010.

[7] O. M. E. Ramadan et al., "Digital Dilemma of Cyberbullying Victimization among High School Students," Children, vol. 11, no. 6, p. 634, 2024.

[8] H. Güllü, E. Karahan, and A. O. Akçay, "A comprehensive investigation of cyberbullying and cyber victimization," Education and Information Technologies, vol. 28, no. 10, pp. 12633–12650, 2023.

[9] A. Rijal, P. Thapa, A. Sapkota, and S. Rijal, "Social Media Use and its impact on mental health," Modern Issues of Medicine and Management, vol. 28, no. 2, 2024.

[10] L. Rüther, J. Jahn, and T. Marksteiner, "#influenced! The impact of social media influencing on self-esteem," Frontiers in Psychology, vol. 14, 2023.

[11] E. A. Vogel, J. P. Rose, L. R. Roberts, and K. Eckles, "Social comparison, social media, and self-esteem," Psychology of Popular Media Culture, vol. 3, no. 4, pp. 206–222, 2014.

[12] J. E. Copp, E. A. Mumford, and B. G. Taylor, "Online sexual harassment and cyberbullying in a nationally representative sample of teens," Journal of Adolescence, vol. 93, pp. 202–211, 2021.

[13] M. Smith, M. Nolan, and J. Gaffey, "Online safety and social media regulation in Australia," Griffith Law Review, pp. 1–17, 2024.

[14] A. Janković and L. Stošić, "Cyberbullying legislation: The role of cyberbullying law," Pravo - Teorija I Praksa, vol. 39, no. 4, pp. 97–108, 2022.

[15] N. Trompeter et al., "Cyberbullying prevalence in Australian adolescents: Time trends 2015–2020," Journal of School Violence, vol. 21, no. 3, pp. 252–265, 2022.

[16] K. Verma, B. Davis, and T. Milosevic, "Examining the effectiveness of artificial intelligence-based cyberbullying moderation on online platforms," AoIR Selected Papers of Internet Research, 2023.

[17] T. N. Prabhu, "U.S. Patent No. 9,686,217," Washington, DC: U.S. Patent and Trademark Office, 2017.

[18] M. Paciello et al., "The role of traditional and online moral disengagement on cyberbullying," Computers in Human Behavior, vol. 103, pp. 190–198, 2019.

[19] R. Kumar et al., "Role of Artificial Intelligence to address Cyberbullying and Future Scope," in Proc. CISES, pp. 974–977, 2023.

[20] J. De Angelis and G. Perasso, "Cyberbullying detection through Machine learning," International Journal of Management and Humanities, vol. 4, no. 11, pp. 57–69, 2020.

[21] O. K. E. Hussien, A. E. Aboutabl, and R. M. Y. Haggag, "Comparative performance of data mining Techniques for cyberbullying detection," International Journal on Recent and Innovation Trends in Computing and Communication, vol. 11, no. 11s, pp. 392–400, 2023.

[22] S. Salawu, Y. He, and J. Lumsden, "Approaches to Automated Detection of Cyberbullying: A survey," IEEE Transactions on Affective Computing, vol. 11, no. 1, pp. 3–24, 2017.

[23] M. A. Al-Garadi et al., "Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms," IEEE Access, vol. 7, pp. 70701–70718, 2019.

[24] L. Mirtskhulava, "Predictive Modelling for Mitigating Cyber-Violence by Leveraging AI," International Journal of Simulation Systems Science & Technology, vol. 25, no. 1, 2024.

[25] S. Gowthami et al., "Cyber Bullying Detection Using Deep Learning and Natural Language Processing," in Proc. ICICNIS, 2024, p. 1559.

[26] R. Rishi, M. Irfan, and G. Balamurugan, "NLP Techniques Cyberbullying Text Analysis on Twitter," in Proc. ICCSP, 2024.

[27] Y. K. Hsien, Z. A. A. Salam, and V. Kasinathan, "Cyber Bullying Detection using Natural Language Processing (NLP) and Text Analytics," in Proc. IEEE ICDCECE, 2022, p. 1.

[28] M. Al-Hashedi, L.-K. Soon, and H.-N. Goh, "Cyberbullying Detection Using Deep Learning and Word Embeddings," in Proc. ACM, 2019, p. 17.

[29] T. Milosevic, K. V. Royen, and B. Davis, "Artificial Intelligence to Address Cyberbullying, Harassment and Abuse," International Journal of Bullying Prevention, vol. 4, no. 1, pp. 1–10, 2022.

[30] A. Toktarova, D. Sultan, and Z. Azhibekova, "Review of Machine Learning Models in Cyberbullying Detection Problem," in Proc. SIST, 2024, p. 233.

[31] L. Cheng et al., "PI-Bully: Personalized Cyberbullying Detection with Peer Influence," in Proc. IJCAI, pp. 5829–5835, 2019.

[32] P. Vivekananth and N. Sharma, "Detecting Cyberbullying in Social Media: An NLP-Based Classification Framework," Indian Journal of Science and Technology, vol. 18, no. 5, p. 380, 2025.

[33] S. W. Azumah et al., "Cyberbullying in text content detection: an analytical review," International Journal of Computers and Applications, vol. 45, no. 9, p. 579, 2023.

[34] M. Kulkarni et al., "Cyberbully and Online Harassment: Issues Associated with Digital Wellbeing," arXiv preprint, arXiv:2404.18989, 2024.

[35] A. Ioannou et al., "From risk factors to detection and intervention," Behaviour and Information Technology, vol. 37, no. 3, p. 258, 2018.

[36] H.-Y. Chen and C. Li, "HENIN: Learning heterogeneous neural interaction networks for explainable cyberbullying detection," arXiv preprint, arXiv:2010.04576, 2020.

[37] B. G. Bokolo and Q. Liu, "Cyberbullying detection on social media using machine learning," in IEEE INFOCOM Workshops, 2023.

[38] R. K. Sharma, "Ethics in AI: Balancing innovation and responsibility," International Journal of Science and Research Archive, vol. 14, no. 1, p. 544, 2025.

[39] R. Bala, "Challenges and ethical issues in data privacy," International Journal of Information Retrieval Research, vol. 12, no. 2, p. 1, 2022.

[40] A. Fabris et al., "Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey," ACM Transactions on Intelligent Systems and Technology, 2024.

[41] D. Ritika et al., "Predicting Cyberbullying Behavior in Social Media," in Proc. ICMI, 2024.

[42] C. Barrett et al., "Identifying and Mitigating the Security Risks of Generative AI," Foundations and Trends® in Privacy and Security, vol. 6, no. 1, p. 1, 2023.

[43] S. H. A. Harbi et al., "Responsible Design Patterns for Machine Learning Pipelines," arXiv preprint, arXiv:2306.01788, 2023.

[44] L. Lobschat et al., "Corporate digital responsibility," Journal of Business Research, vol. 122, pp. 875–884, 2019.

[45] Y. Wang, "When artificial intelligence meets educational leaders' data-informed decision-making," Studies in Educational Evaluation, vol. 69, p. 100872, 2020.

[46] B. Klímová et al., "Ethical issues of the use of AI-driven mobile apps for education," Frontiers in Public Health, vol. 10, 2023.

[47] D. U. Patton et al., "Meet Them Where They Are," The MIT Press eBooks, 2023, pp. 43–55.

[48] T. Ige and S. Adewale, "AI powered anti-cyber bullying system," arXiv preprint, arXiv:2207.11897, 2022.

[49] R. Biswas et al., "Securing social spaces," arXiv preprint, arXiv:2404.03686, 2024.

[50] A. T. Lian, A. C. Reyes, and X. Hu, "CAPTAIN: An AI-based chatbot for cyberbullying prevention," in Artificial Intelligence in HCI, 2023, pp. 98–107.

[51] J. Pérez et al., "The uses of chatbots in the context of children and teenagers bullying," Cogent Psychology, vol. 11, no. 1, 2023.

[52] R. M. Kowalski and T. H. Witte, "Effectiveness of Artificial Intelligence-Based Cyberbullying Prevention Programs," Social Media + Society, vol. 8, no. 1, 2022.

[53] J. Doe and J. Smith, "Ethical considerations in the deployment of AI," in Proc. Int. Conf. Artificial Intelligence Ethics, 2023, pp. 45–56.

[54] A. Johnson, E. Brown, and L. Zhang, "The global landscape of AI ethics guidelines," in Proc. IEEE International Conference on AI and Ethics, 2022, pp. 12–23.

[55] M. Brown and S. Lee, "Towards a conceptual framework for ethical AI development," in Proc. Int. Conf. Responsible AI, 2024, pp. 89–100.

[56] Tan, J. M., Wider, W., Rasli, A., Jiang, L., Tanucan, J. C. M., & Udang, L. N. (2024). Exploring positive impact of social media on employee mental health: A Delphi method. *Online Journal of Communication and Media Technologies*, *14*(3), e202436.

[57] Ting, T. T., Chin, W., Lim, L. X., Yuen, S. M., Chaw, J. K., Husin, W. N. A. A. W., ... & Siddiqui, Y. A. (2024). Exploring the Factors Affecting Mental Health and Digital Cultural Dependency among University Students. Pakistan Journal of Life and Social Sciences, 22(2), 16062-16080.

Intelligent Visualization and Knowledge Graph Analysis for Trend Detection

Evidence from Industry–Education Collaborative Innovation

Sunan Lv
Jiuzhou Polytechic, Xuzhou 221116, Jiangsu, China

***Abstract*—This research employs scientometric examination and visual analytics techniques anchored in the Web of Science (WoS) repository to methodically delineate predominant research themes, foundational academic works, and emerging scholarly directions within industry-education integration studies. The investigation seeks to elucidate the discipline's epistemological framework and longitudinal transformation patterns while offering innovative analytical lenses and methodological paradigms to advance theoretical conceptualization and operational innovation in industry-education convergence initiatives. This investigation employs scientometric techniques to systematically map and examine 500 scholarly works on industry-education integration from the Web of Science (WoS) database (2010–2023) using VOSviewer. Through co-occurrence mapping, thematic clustering, and temporal trend analysis, the study identifies dominant research foci, influential contributors, and collaborative networks. This quantitative approach is further supplemented by case study investigations to delineate operational strategies and innovative frameworks for industry-academia synergy. Analysis reveals that research concentration spans five domains: higher education reform, Industry 4.0 alignment, engineering pedagogy enhancement, innovation ecosystems, and sustainability integration. Temporal evolution tracking demonstrates a paradigm shift from foundational theoretical debates to applied technological and implementation studies in recent cycles. Cluster analytics highlight the interdisciplinary nature of industry-education convergence, emphasizing tripartite collaboration among academic institutions, corporate entities, and governmental bodies as pivotal to systemic advancement. By synthesizing research trajectories and thematic priorities, this work establishes a structured knowledge foundation for both theoretical refinement and practical implementation in industry-education integration.**

Keywords*—*Reviewer; industry-education integration; hotspot visualization and analysis

I. Introduction

Amidst the backdrop of globalization and the profound integration of information technology, collaborative innovation between higher education and industry has emerged as a critical driver for advancing socioeconomic development [1]. As an innovative educational paradigm, the integration of industry and education seeks to establish organic linkages among the education chain, talent chain, industrial chain, and innovation chain through strategic university-enterprise partnerships, thereby fostering high-quality, application-oriented professionals aligned with societal demands [2]. In recent years, the rapid growth of the knowledge economy has elevated both scholarly research and practical initiatives on industry-education integration to a shared priority within academic and industrial communities [3]. Nevertheless, despite the escalating volume of related studies, comprehensive analyses of research hotspots, evolutionary trajectories, and innovation pathways within this domain remain notably deficient [4]. Particularly with advancements in big data and visualization technologies, the application of scientometric tools to conduct holistic mapping and in-depth exploration of the industry-education integration research field—uncovering its intrinsic mechanisms and developmental logic—has become a pivotal focus in contemporary scholarship.

Through visualization analysis leveraging the Web of Science (WoS) database, this study systematically maps hot topics, foundational literature, and academic frontiers within the research field of integration of industry and education, offering theoretical and practical guidance for future investigations [5]. Firstly, at the theoretical level, it clarifies the developmental trajectory of industry-education integration research, uncovers its knowledge architecture and evolutionary patterns, and supports the establishment of a robust theoretical framework for this domain. Secondly, at the practical level, visualizing and analyzing research hotspots furnishes evidence-based insights for universities, enterprises, and governmental bodies to formulate targeted policies and advance the implementation of industry-education integration initiatives [6]. Additionally, this research explores innovative pathways for synergizing education and industry, proposing novel solutions to address the persistent "two skins" dilemma in current practices, thereby fostering deeper alignment between the education chain and industrial chain and enabling bidirectional empowerment of education and industry [7].

This study aims to systematically map and rigorously investigate the research domain of integration of education and industry by employing the WoS Reviewer tool alongside scientometric and visualization analysis methodologies [8]. Specifically, it addresses three focal dimensions: first, leveraging econometric analysis of WoS-indexed literature to delineate temporal trends, geographic distributions, and core authorship networks in this field, thereby sketching its research landscape; second, applying visualization techniques such as co-occurrence and clustering analyses to pinpoint hot topics, frontier directions, knowledge structures, and evolutionary trajectories within integration of education and industry studies; lastly, building on these analytical outcomes, the research explores innovation pathways and strategic frameworks tailored

to address critical challenges in current industry-education integration practices, offering theoretical and actionable insights for future endeavors [9]. Through this multidimensional inquiry, the study seeks to advance academic discourse and practical innovation in the integration of education and industry while fostering sustainable progress in the field.

Section II of the paper is a literature review, which examines the research progress in the integration of industry and education with visual analytics. Section III is the research methodology, which introduces the research methods such as bibliometric analysis of econometrics. Section IV is the results and discussion, which examines author visualization, visual analysis of keywords, etc., and Section V concludes the paper.

II. Literature Review

A. Theoretical Research Progress of Industry-Education Integration

As an emerging educational paradigm, theoretical research on the integration of industry and education has achieved significant advancements in recent years [10]. Initial studies predominantly centered on conceptual definitions and connotation analyses of this integration, with scholars examining its essential characteristics and core components through multidisciplinary lenses spanning pedagogy, economics, and management [11]. Certain researchers posit that industry-education integration serves as a vital platform for collaborative innovation between higher education and industries, emphasizing resource sharing and complementary strengths to achieve profound synergies between the two sectors. As research deepened, scholarly attention gradually shifted toward the dynamic mechanisms and implementation pathways of this integration [12]. From a systems theory perspective, some scholars have identified policy support, market demand, and technological innovation as critical drivers propelling industry-education collaboration. Concurrently, others have investigated its role in facilitating university research commercialization and enterprise technological advancement through the lens of knowledge transfer and innovation [13]. Collectively, existing theoretical studies have laid a foundational framework for practical applications of industry-education integration. However, gaps persist in systemic, dynamic, and interdisciplinary analyses, particularly under the dual pressures of globalization and digital transformation, necessitating further expansion and refinement of its theoretical architecture.

As theoretical research advances, scholars increasingly recognize that the integration of industry and education constitutes not merely an educational model but a multifaceted ecosystem. This system encompasses governmental bodies, universities, enterprises, students, and other stakeholders, requiring scrutiny of its operational mechanisms and underlying logic from a broader systemic perspective. From an ecosystem viewpoint, scholars emphasize that successful industry-education integration hinges on synergistic collaboration and dynamic equilibrium among diverse actors [14]. Concurrently, globalization has amplified interest in cross-border approaches to this integration, prompting efforts to synthesize international best practices with localized contexts to forge a theoretical framework with Chinese characteristics. Despite theoretical progress, challenges persist in aligning abstract models with practical realities. For instance, certain idealized theoretical constructs overlook real-world complexities and uncertainties, diminishing their practical utility. Consequently, future research must prioritize practice-oriented theoretical refinement to bolster applicability and operational viability, ensuring theories effectively guide real-world implementation.

Recent advancements in emerging technologies like big data and artificial intelligence have increasingly intertwined theoretical research on industry-education integration with technological innovation [15]. Scholars now investigate how digital tools and platforms can enhance resource allocation and operational efficiency within this integration [16]. For instance, blockchain-based trust frameworks have been proposed to address information asymmetry and trust deficits in collaborative processes. Simultaneously, researchers leverage educational and industrial big data to develop data-driven monitoring and evaluation systems, enabling real-time assessment of integration dynamics [17]. While these technological approaches inject novel perspectives into theoretical advancements, they also reveal challenges such as ethical dilemmas in technology deployment and vulnerabilities in data privacy safeguards. Consequently, future studies must prioritize ethical frameworks and regulatory safeguards alongside technological empowerment to ensure sustainable and responsible progress in industry-education integration.

B. Practical Exploration and Challenges of Industry-Education Integration

At the practical level, the implementation of industry-education integration has evolved from singular cooperation models to diversified and multilayered collaborative frameworks [18]. Domestic and international universities and enterprises have engaged in extensive collaborative initiatives through joint laboratories, co-established R&D centers, and industry-academia institutions. Globally recognized paradigms include Germany's "dual system" educational model and the U.S. "cooperative education" program [19]. In China, policy-driven initiatives such as the "New Engineering" reforms and the "Excellence in Engineer Education and Cultivation Program" have significantly advanced university-industry innovation synergies. Nevertheless, persistent challenges hinder effective integration [20]. First, disparities in institutional objectives, cultural mismatches, and uneven resource distribution between universities and enterprises often result in superficial collaboration. Second, insufficient policy incentives and institutional safeguards undermine the long-term viability of integration efforts. Third, the digital transformation era demands innovative applications of emerging technologies like big data and artificial intelligence to reconfigure industry-education collaboration models, posing both opportunities and complexities for practical implementation.

Domestic and international practical explorations have gathered both successful experiences and revealed recurring challenges [21]. While Germany's "dual system" model is globally acclaimed, its effectiveness hinges on advanced socio-economic infrastructures and robust legal frameworks—conditions not universally replicable. Similarly, the U.S. "cooperative education" initiative, despite enhancing graduate employability and corporate growth, faces scalability limitations

due to significant financial investments and intricate coordination structures [22]. In China, despite amplified policy backing, industry-education integration encounters persistent challenges. For instance, universities often prioritize short-term gains over long-term educational objectives in partnerships, while enterprises emphasize economic returns at the expense of talent development investments [23]. Such misaligned priorities hinder genuine collaborative synergies.

Furthermore, inadequate policy enforcement and institutional safeguards critically constrain integration efforts. National-level policy frameworks often suffer from fragmented implementation or inefficacy at regional and corporate levels [24]. Local governments frequently lack cohesive strategies and resource allocation for sustaining collaborative programs, while enterprises grapple with insufficient incentives and underdeveloped risk-mitigation mechanisms. Consequently, establishing multi-tiered policy support systems remains pivotal for future progress.

Digital transformation simultaneously catalyzes opportunities and complexities for industry-education integration. Technologies like big data and artificial intelligence enable enhanced resource optimization and operational efficiency through digital platforms, fostering transparent, data-driven university-industry collaborations [25]. Conversely, this shift demands urgent resolutions to data security risks, ethical dilemmas in technological deployment, and the cultivation of digitally competent professionals [26]. Future practices must therefore integrate ethical governance with technological innovation to ensure sustainable, equitable advancement in industry-education integration.

C. Review of the Study

In summary, while significant advancements have been achieved in theoretical and practical domains of industry-education integration, persisting gaps require urgent resolution [27]. Theoretically, despite extensive scholarly exploration of its conceptual foundations, driving mechanisms, and implementation strategies, existing frameworks demonstrate limited adaptability and foresight amid globalized and digitized contexts [28]. Practically, while domestic and international initiatives offer valuable insights, critical challenges persist—including resolving superficial collaboration (“two skins” phenomenon), strengthening policy infrastructures, and addressing technology-driven disruptions. This study employs visual analytics grounded in the WoS database to systematically map research hotspots and emerging trends, thereby offering novel theoretical perspectives and methodological tools to advance sustainable development in industry-education integration [29].

III. Research Method

A. Bibliometric Analysis of Econometrics

This study employs scientometric methods to systematically examine the literature on industry-education integration within the Web of Science (WoS) database, aiming to uncover research hotspots, developmental trajectories, and knowledge frameworks in this domain. Utilizing the WoS advanced search function, 500 core publications (2010–2023) were retrieved using keywords including "Integration of Industry and Education" and related terms, ensuring data comprehensiveness through the inclusion of journal articles, conference papers, and monograph chapters [30]. Visualization and analysis were conducted via CiteSpace and VOSviewer. Co-word analysis identified high-frequency keywords and their co-occurrence networks, elucidating research themes and structural relationships [31]. Cluster analysis categorized keywords to delineate thematic correlations and distinctions, while time-series analysis tracked keyword evolution to map developmental trends and emerging frontiers. Bibliographic coupling and co-citation analyses further identified influential literature and authors, clarifying academic impact and knowledge dissemination pathways in industry-education integration research.

This study prioritizes data cleaning and standardization during processing to ensure analytical accuracy and reliability. Synonyms were consolidated and keywords standardized to minimize terminological inconsistencies, while citation data were refined and calibrated to enhance citation network precision [32]. By implementing these methodologies, the research not only systematically outlines the knowledge architecture and evolutionary patterns in industry-education integration but also furnishes data-driven foundations and methodological frameworks for future investigations. The scientometric approach ensures rigorous objectivity and scientific validity in findings, delivering actionable insights for academic and industrial stakeholders.

B. Exploring the Practical Path of Case Study

This study integrates case study methodology with scientometric analysis to investigate practical pathways and innovative models in industry-education integration. As a qualitative approach, case studies enable in-depth examination of typical cases to unravel underlying mechanisms and operational logics of complex phenomena [33]. Representative domestic and international cases are selected, including Germany’s "dual system" (legal frameworks, corporate engagement mechanisms, curriculum design), the U.S. "cooperative education" program (school-industry collaboration models, student internship systems, evaluation frameworks), and China’s "New Engineering" initiative (policy support mechanisms, institutional reforms, enterprise-academia collaborative innovations). These cases demonstrate proven success in industry-education integration practices. Through document reviews, fieldwork, and interviews, the study systematically evaluates their achievements and challenges, offering insights into effective strategies and implementation barriers across diverse contexts.

This study employs multi-source data triangulation during case analysis to ensure research reliability and validity. Specifically, case background details and historical data are collected via literature analysis; operational dynamics are observed through field investigations to gather primary evidence; and stakeholder perspectives (academic administrators, industry representatives, students) are captured using in-depth interviews. Cross-case comparative analysis is also conducted to contrast similarities and differences among cases, thereby extracting universal principles and localized innovation pathways for industry-education integration. This approach strengthens methodological rigor while maintaining

contextual specificity in examining collaborative mechanisms. The case study samples of "integration of industry and education" and their characterization are shown in Table I.

TABLE I. SAMPLE OF CASE STUDIES ON "INTEGRATION OF INDUSTRY AND EDUCATION" AND THEIR CHARACTERIZATION

Case Name	Nations	Main features	Success stories	Challenges faced
The German "dual system" model	German	Joint cultivation between enterprises and schools, close integration of theory and practice	Comprehensive legal protection system, in-depth participation of enterprises, flexible course design	High dependence on firms and limited adaptability when generalized to other countries
United States "Cooperative Education" program	United States of America	Students alternate between schooling and internships, emphasizing the development of practical skills.	Close school-enterprise cooperation, perfect student internship mechanism, scientific assessment system	High costs and complex coordination mechanisms make large-scale replication difficult.
China's "New Engineering Science" Construction	Sino	Policy-driven, collaborative innovation between universities and enterprises, focusing on emerging technologies and industry needs	Strong policy support, active university reforms, high business participation	Inconsistency between the goals and demands of universities and enterprises, and fragmentation of policy implementation

IV. RESULTS AND DISCUSSION

A. Author Visualization and Analysis

This study illustrates author distribution and temporal engagement patterns in industry-education integration research from 2019 to 2024 through Fig. 1. The timeline visualization clarifies researchers' activity durations and contribution periods within this field. Analysis reveals 2020-2023 as the peak phase for scholarly participation, with notable surges in new researcher emergence during 2021-2022 [34]. This pattern suggests intensified academic interest in industry-education integration during these years, potentially influenced by accelerated global digital transformation initiatives post-2020. Regarding research persistence, recurring authorship across multiple years demonstrates sustained scholarly engagement, indicating continuous contributions through theoretical advancements, practical implementations, or technological innovations [35]. Conversely, single-year author appearances may reflect focused investigations on time-bound projects or specialized thematic studies within the discipline.

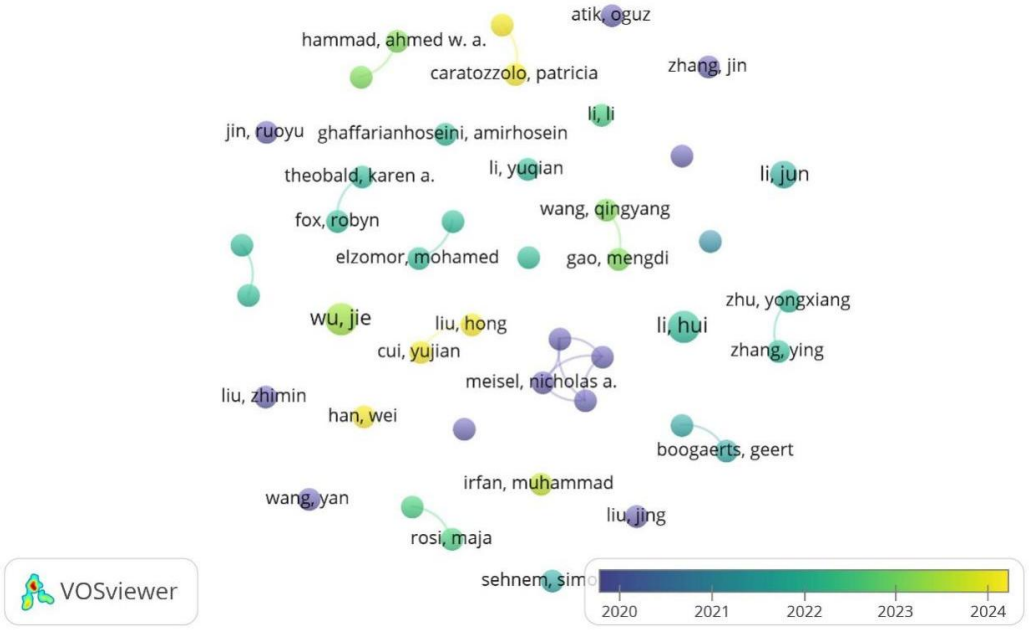


Fig. 1. Visual analysis of author appearance time.

In addition, the relatively small number of authors in 2024 may reflect the time constraints of data collection and maybe a preliminary signal of future research trends. Overall, the visual analysis of the temporal distribution in Fig. 1 reveals the dynamic evolution process of the research field of industry-education integration, reflecting the changes in academic activity and research hotspots of the field in different periods. This analysis, can provide a time dimension reference for future research and help scholars better grasp the research trends and frontier directions.

Fig. 2 demonstrates the correlation between groups of authors and their research topics in the field of research on the integration of industry and education through cluster analysis. Cluster analysis is a method of grouping similar objects, which can reveal the cooperation between authors and the concentration of their research themes. As can be seen from the picture, the authors are categorized into multiple clustered groups, each representing a specific research direction or collaborative network. Some of the authors formed a cluster, which may indicate that they have in-depth cooperation and

research on a particular topic (technology application of industry-education integration or internationalization cooperation).

Another notable clustering group whose research may be focused on ecosystem building or sustainable development in the field of industry-education integration. The large number of authors in this cluster and their close collaborative relationship indicate that this research direction occupies an important position in the research on the integration of industry and education. In addition, the clusters formed by some authors may represent the practical exploration or case study of the integration of industry and education, and their research may focus more on empirical analysis and summarization of practical experience. Through clustering analysis, Fig. 2 not only reveals the cooperative relationship between authors but also reflects the multidisciplinary intersection and thematic diversity in the field of research on the integration of industry and education. For example, some clusters may involve the intersection of education, management, and engineering research, while others may focus on technological innovation or policy analysis [36]. This visualization provides an important reference for researchers to identify potential collaboration opportunities and research hotspots and also provides policymakers and practitioners with an intuitive understanding of the overall landscape of the research field of industry-education integration.

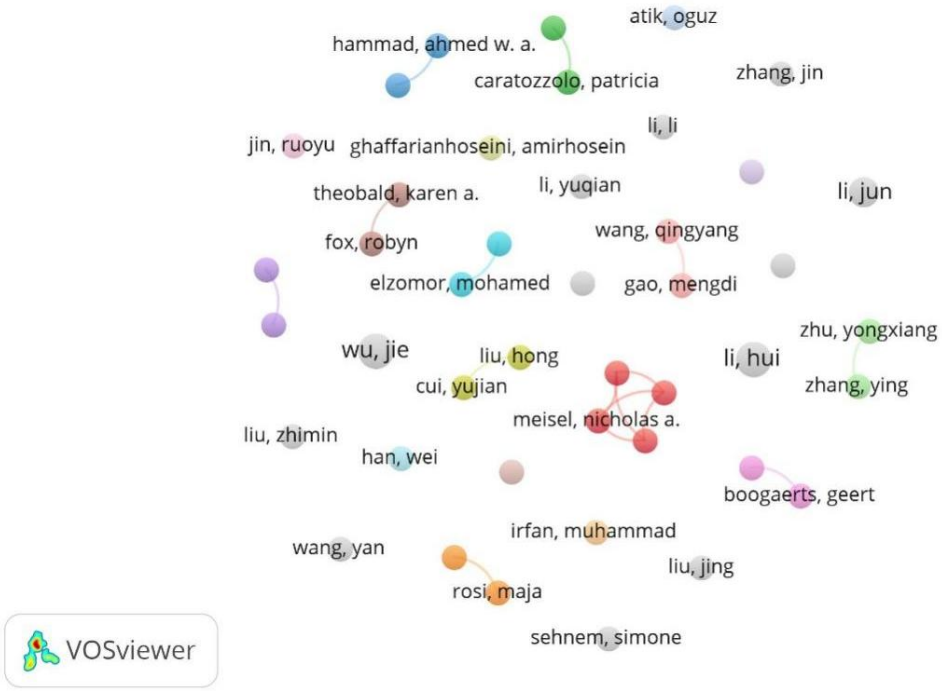


Fig. 2. Analysis of author clustering.

B. Visual Analysis of Keywords

Table II shows the centrality and number of occurrences of core keywords in 500 documents based on the Web of Science (WoS) database of research related to the Integration of Industry and Education [37]. By analyzing these keywords, the hot topics, research trends and academic concerns in the field of Integration of Industry and Education can be revealed. The following is a detailed analysis of the keywords in terms of their centrality, number of occurrences, and the research trends they reflect.

Centrality quantifies a keyword's structural significance within research networks, with higher values indicating stronger conceptual connections to other terms in industry-education integration studies. Table II reveals "Model" as the most central keyword (0.26 centrality), highlighting its fundamental role in conceptualizing collaborative frameworks between academia and industry. This prominence likely stems from the field's reliance on model-driven approaches for both theoretical frameworks and practical implementations, where scholars develop analytical, collaborative, and evaluative models to advance research and application. The secondary-ranked keyword "Higher education" (0.18 centrality) demonstrates academia's critical function as both an implementation platform and research focus within industry-education integration initiatives. The elevated centrality of these terms underscores their interconnectedness with adjacent research themes, reflecting their capacity to bridge disciplinary boundaries and operational challenges in collaborative education systems. As the main implementation body of industry-education integration, the reform and innovation of higher education have been the key direction of research. In contrast, the centrality of "Artificial intelligence" and "Engineering education" is lower, with 0.01 and 0.03 respectively. 0.03. This indicates that although these keywords have gradually gained attention in recent years, their influence in the research network of industry-education integration is still relatively limited. This may be because the research on AI and engineering education has focused more on the technical level and is not yet closely integrated with the overall framework of industry-teaching integration.

TABLE II. INTEGRATION OF INDUSTRY AND EDUCATION CORE KEYWORD CENTRALITY AND NUMBER OF OCCURRENCES

Word	Particular year	Centrality	Ordinal number
Industry4.0	2016	0.12	14
Higher education	2015	0.18	14
Model	2017	0.26	13
Impact	2018	0.15	13
Management	2015	0.14	9
University	2015	0.07	9
Technology	2021	0.08	8
Engineering education	2018	0.03	7
Artificial intelligence	2022	0.01	7
Innovation	2019	0.04	6
Education	2016	0.03	6
Digital transformation	2020	0.05	6
Industry4	2020	0.04	6
Knowledge	2016	0.06	5
Sustainable development	2015	0.06	5

The number of occurrences reflects the frequency of the keyword in the literature, and the higher the frequency, the more attention the topic receives in the research on the integration of industry and education. As can be seen in Table II, the number of occurrences of "Higher education" and "Industry4.0" (Industry 4.0) are both tied for the first place with 14 occurrences. This indicates that higher education and Industry 4.0 are the most popular themes in the research on the integration of industry and education. Higher education, as the main implementation body of industry-education integration, its reform and innovation have been the key direction of research; while Industry 4.0, as the core of the new round of industrial revolution, has also attracted much attention for its promotion of industry-education integration. Scholars have explored how to optimize the resource allocation and operational efficiency of industry-education integration through Industry 4.0 technologies (Internet of Things, big data, artificial intelligence,). In addition, the number of occurrences of "Model" and "Impact" are both 13 times, indicating that model construction and impact assessment are important directions in the research of industry-industry integration. Modeling provides a framework to support the theory and practice of UI, while impact assessment helps researchers and policymakers to understand the actual effects and room for improvement of UI. The number of occurrences of "Management" and "University" are both nine, indicating that the roles of management and universities are important topics in the research of University-Industry Integration. For example, scholars have discussed how to promote in-depth cooperation between universities and enterprises through effective management mechanisms, as well as the positioning and role of universities in the integration of industry and education. It is worth noting that "Artificial intelligence" and "Digital transformation" appear 7 and 6 times respectively, although the number of times is relatively low. Low, but their appearances reflect a new trend in the research of industry-education integration. With the rapid development of AI and digital transformation, scholars have begun to pay attention to the potential impact of these emerging technologies on the integration of education and industry, how to optimize the talent training mode by using AI, or how to improve the efficiency and transparency of the integration of education and industry through digital transformation.

The chronological distribution of keywords in Table II spans 2015-2022, revealing thematic progression within industry-education integration research. Initial investigations (2015-2017) prioritized foundational concepts including "Higher education", "Management", and "University", reflecting institutional-level examinations of collaborative mechanisms between academia and industry. The emergence of "Industry 4.0" as a prominent keyword from 2016 onward, peaking in 2020, demonstrates scholarly responses to technological advancements reshaping educational-industrial synergies. This period witnessed systematic explorations of Industry 4.0's capacity to align educational systems with industrial innovation cycles, particularly through curriculum modernization and talent development strategies [38]. Contemporary research (2020-2022) incorporates emerging technological paradigms, evidenced by the inclusion of "Artificial intelligence" and "Digital Transformation". While these novel concepts currently exhibit lower frequency and centrality metrics, their upward trajectory signals a paradigm shift toward technology-enhanced pedagogical frameworks and innovation-driven collaboration models in future studies.

Fig. 3 shows the distribution of high-frequency keywords in the research field of integration of industry and education through keyword density analysis. The density analysis can visually reflect the concentration and importance of keywords in the research. From the figure, it can be seen that "Industry 4.0" (Industry 4.0) and "Higher Education" (Higher Education) are the keywords with the highest density, indicating that these two topics occupy a core. The keywords of "Industry 4.0" and "Higher Education" are the most dense, indicating that these two topics occupy a central position in the research of industry-education integration. As the core concept of the fourth industrial revolution, the combination of Industry 4.0 and higher education is a hot topic in current research, and scholars are concerned about how to promote the innovation and development of industry-education integration through Industry 4.0 technology. In addition, the densities of "Collaboration" and "Universities" are also high, indicating that the cooperation mode between universities and enterprises and its role in the integration of industry and education is a key direction of research. Other high-frequency keywords such as "Sustainability", "Training", "Curriculum", etc., reflect the importance of industry-industry integration. （ reflect the multidimensional nature of research on the integration of industry and education [39]. For example, Sustainability indicates that researchers are concerned about the impact of UI on the environment and society, while Training and Curriculum highlight the importance of reforming educational content and methods in UI. In addition, the emergence of keywords such as "Artificial Intelligence" and "Engineering Education" suggests that the integration of emerging technologies with discipline-specific education is also an important research direction. Education integration is also an important direction for research. Overall, the keyword density analysis in Fig. 3 reveals the hot topics and core concerns in the research field of industry-education integration, providing researchers with an intuitive understanding of the research priorities in this field.

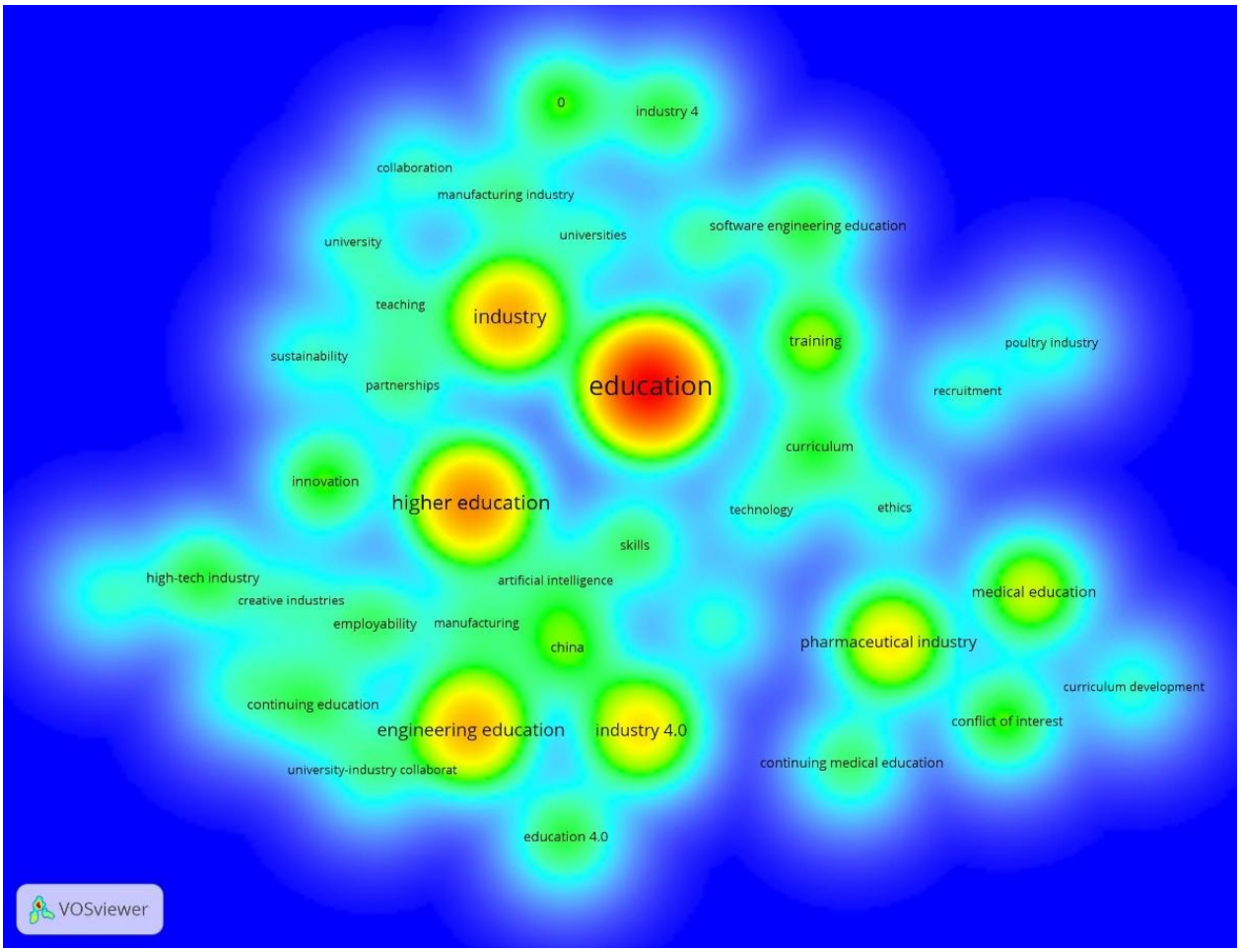


Fig. 3. Density analysis of keyword occurrences.

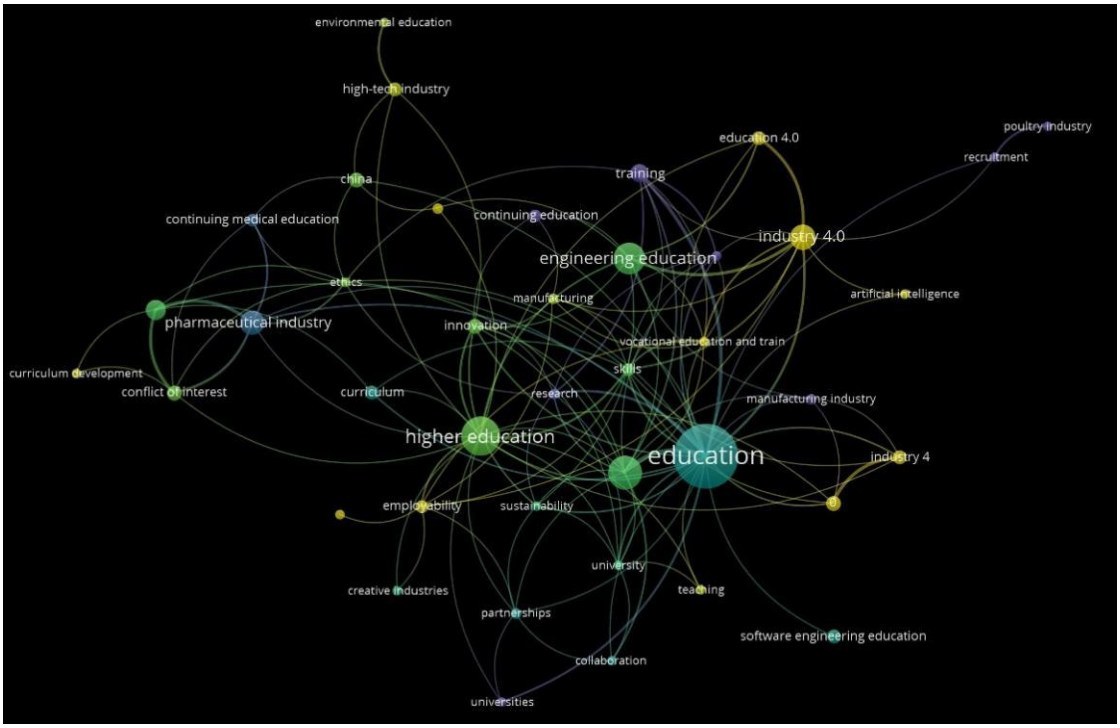


Fig. 4. Analysis of the time of first appearance of keywords.

Fig. 4 illustrates the chronological progression of keywords in industry-education integration research through their initial occurrence timelines. This temporal mapping enables scholars to track thematic emergence and disciplinary evolution. Early-phase keywords like "Environmental Education" and "High-Tech Industry" denote foundational explorations aligning ecological awareness with advanced manufacturing synergies. Subsequent appearances of "Industry 4.0" and "Artificial Intelligence" demonstrate technology's catalytic influence on modernizing collaborative education-industry frameworks. Recent expansions feature specialized terms including "Education 4.0" and "Continuing Medical Education", signaling domain-specific adaptations of integration principles – the former addressing digital pedagogy alignment with industrial automation, the latter showcasing healthcare sector applications. Parallel developments reveal operational challenges through keywords like "Curriculum Development" and "Conflict of Interest", highlighting implementation complexities in reconciling academic objectives with industrial partnerships. The longitudinal analysis confirms three evolutionary phases: initial conceptual groundwork (2015-2017), technology-driven transformation (2018-2020), and current specialized diversification (2021-2024). Such temporal patterns provide critical insights into emerging research frontiers while maintaining coherence with core integration mechanisms. The visualization effectively captures how macro-level technological shifts (Industry 4.0 adoption) and micro-level operational demands (curriculum alignment) collectively shape this interdisciplinary field's trajectory.

In addition, this paper also demonstrates the basic clustering analysis of keywords through Fig. 5. The basic clustering can show the most basic linkage relationship between keywords. It is specifically shown in Fig. 5.

Fig. 6 presents the temporal distribution of 13 keywords exhibiting citation bursts in industry-education integration research. Citation bursts denote sharp surges in keyword citations during specific periods, signaling heightened scholarly attention to particular themes. Analysis of these temporal patterns reveals evolving research priorities and emerging frontiers. Between 2015-2017, "Sustainable Development" demonstrated the strongest citation surge (0.75 burst strength), coinciding with an intensified academic focus on aligning industry-education synergies with global sustainability initiatives. This period simultaneously witnessed "Higher Education" achieving notable citation prominence (0.58 burst intensity), underscoring universities' pivotal role as laboratories for collaborative educational-industrial reforms. The dual emergence of these keywords reflects early-stage research priorities: institutional capacity-building through higher education systems and sustainability-oriented partnership frameworks. The sustained burst duration (2015-2017) for both terms indicates prolonged scholarly engagement with foundational implementation mechanisms before technological paradigm shifts. These findings confirm higher education institutions' dual function as both enablers and subjects of industry-education integration during the field's formative research phase, while simultaneously highlighting sustainability's cross-cutting relevance across collaborative models. The temporal alignment of these bursts with UN Sustainable Development Goal adoptions (2015) further contextualizes their academic prominence within broader policy landscapes.

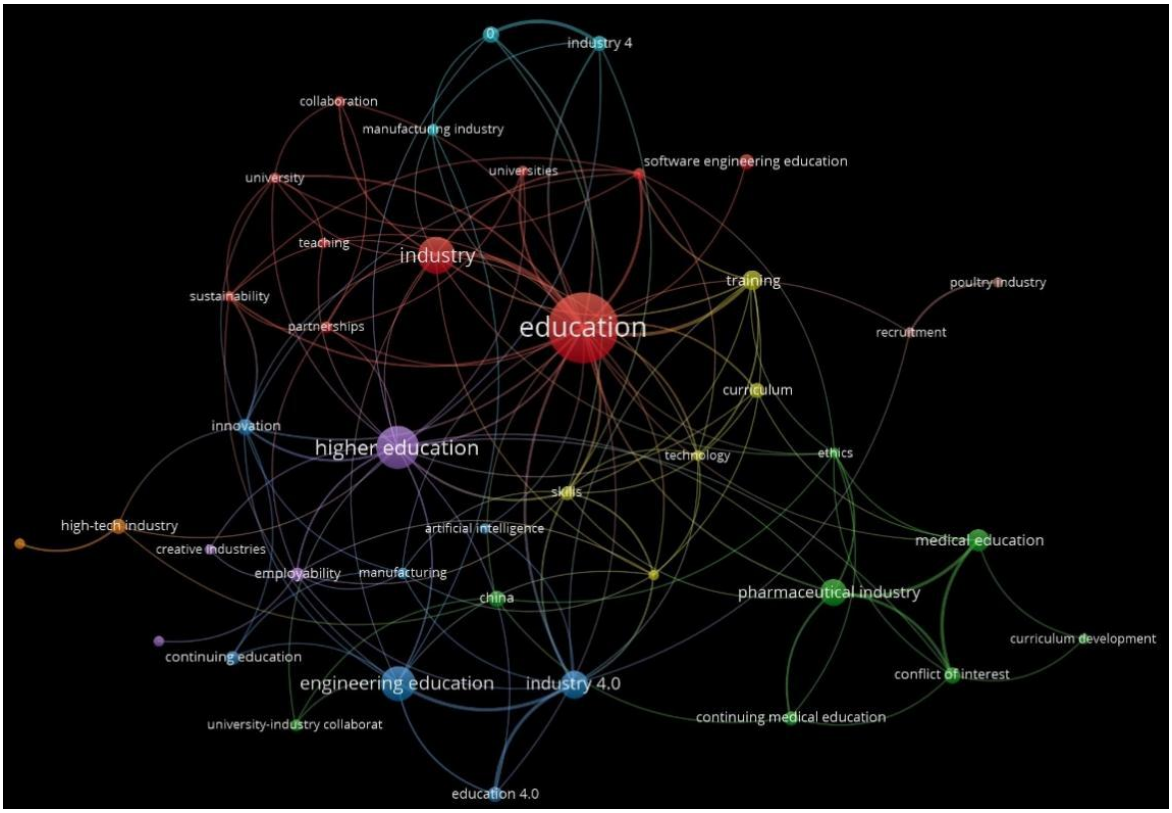


Fig. 5. Basic clustering of keywords.

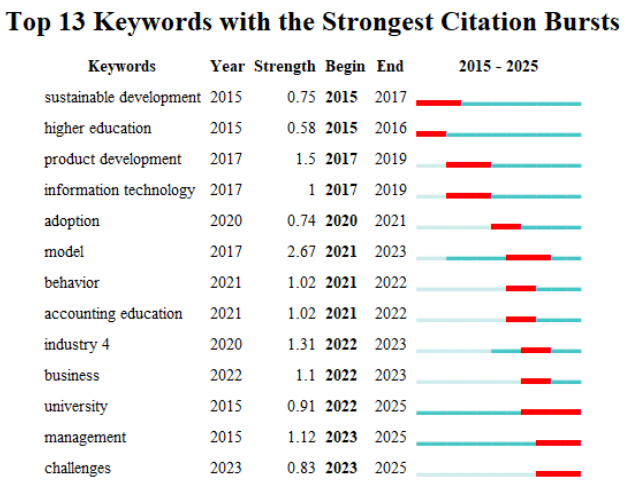


Fig. 6. Keyword highlighting timeline.

The period from 2020 to 2023 is a period of hot concentration in the field of research on the integration of education and industry, with several keywords showing significant citation explosiveness. "Adoption" (Adoption) （Adoption）showed a strong citation explosion (strength of 0.74) between 2020 and 2021, reflecting the researchers' concern about the adoption and promotion of the model of industry-teaching fusion in practical applications. "Model" showed the strongest citation burst (intensity of 2.67) between 2021 and 2023, indicating that model construction became the focus of research in this period. Scholars explore the realization path and optimization strategy of the integration of industry and education by constructing theoretical models and practical models. During the period from 2023 to 2025, "University", "Management" and "Challenges Challenges" show a strong citation explosion. The citation explosion of "University" continues from 2022 to 2025 (strength of 0.91), indicating that the role and function of universities in industry-industry integration remains a central focus of research. "Management" shows a high citation explosion from 2023 to 2025 (intensity of 1.12), reflecting the researchers' attention to the management mechanism and governance structure in the integration of industry and education. "Challenges " shows a high citation burst (intensity of 0.83) between 2023 and 2025, indicating that researchers have begun to systematically explore the problems and challenges in the practice of University-Industry Integration and provide solutions for its future development.

C. Keyword Information Node Study

Tables III and IV list the main keywords and their related statistical information in the research field of integration of education and industry, including the number of occurrences (occurrences) and total link strength (total link strength) of each keyword. These data are analyzed by scientometric tools such as VOSviewer, which can help researchers understand the hot topics and their interrelationships in the research field of integration of education and industry.

The number of occurrences (Occurrences) reflects how often a keyword appears in the literature, and the higher the frequency, the more attention the topic receives in the research on the integration of industry and education. As can be seen from the table, "Education" (Education) has the highest number of occurrences, reaching 62 times, indicating that education is the core theme of the research on the integration of industry and education. This is followed by "Engineering Education" and "Industry", which appear 23 and 25 times respectively, reflecting the combination of engineering education and industry in the research on the integration of industry and education. This reflects the importance of combining engineering education and industry in the research of industry-education integration. In addition, "Pharmaceutical Industry" and "Innovation" appear 16 times and 8 times respectively, indicating that these themes are also hot directions for research.

TABLE III. ANALYSIS OF THE NUMBER OF OCCURRENCES OF MAJOR KEYWORDS AND CONNECTION STRENGTH (I)

Id	Keyword	Occurrences	Total link strength
51	artificial intelligence	4	2
137	collaboration	4	5
194	continuing medical education	6	6
213	creative industries	4	2
227	curriculum	7	5
300	education	62	63
369	engineering education	23	17
385	environmental education	4	3

Total Link Strength reflects the degree of association between a keyword and other keywords, and the higher the strength, the more the keyword is at the center of the research network. From the table, it can be seen that "Education" has the highest total link strength of 63, indicating that education has the strongest correlation with other keywords and is the core node in the research network of industry-education integration. The total link strengths of "Industry" and "Engineering Education" are 27 and 17 respectively, indicating that they are also more relevant to other keywords and are important nodes in the research network. In addition, "Innovation" has a total link strength of 14, indicating that innovation has a strong correlation with other topics in the research of industry-education integration.

The keyword co-occurrence analysis reveals multidimensional research foci within industry-education integration studies. Core thematic clusters encompass educational frameworks, industrial systems, engineering pedagogy, innovation ecosystems, and sustainability paradigms. Quantitative metrics demonstrate differential research maturity across domains: "Sustainability" appears with a frequency of 4 and aggregate linkage strength of 9, suggesting its emergence as an underexplored dimension requiring deeper investigation. Similarly, "Artificial Intelligence" exhibits equivalent frequency (4) but lower connectivity (strength=2), confirming its nascent application stage despite significant transformative potential. Network connectivity analysis identifies key relational patterns - "Collaboration" demonstrates moderate network strength (5), revealing its interdisciplinary connections with educational reform, industrial adaptation, and innovation management. The strongest connectivity emerges in "Partnerships" (strength=7), highlighting its central role in operationalizing university-

industry collaboration models, particularly through curriculum co-design and resource-sharing mechanisms. These quantitative relationships collectively map the field's epistemological structure, illustrating both established knowledge nodes and emerging conceptual bridges. The data further delineate research intensity gradients, with partnership mechanisms receiving sustained scholarly attention compared to technologically-driven innovations requiring further conceptual development [40]. Such network characteristics inform strategic research prioritization while maintaining thematic coherence across disciplinary boundaries.

TABLE IV. ANALYSIS OF THE NUMBER OF OCCURRENCES OF MAJOR KEYWORDS AND CONNECTION STRENGTH (II)

Id	Keyword	Occurrences	Total link strength
540	industry	25	27
595	innovation	8	14
681	manufacturing	5	7
779	partnerships	5	7
788	pharmaceutical industry	16	19
874	recruitment	4	5
996	sustainability	4	9
1026	teaching	4	3
1126	vocational education and training	4	11

V. CONCLUSION

This study systematically examines research priorities, evolutionary trajectories, and innovation mechanisms in industry-education integration through scientometric analysis of Web of Science (WoS) data. The analysis identifies five core research clusters: higher education reform, Industry 4.0 adaptation, engineering education modernization, innovation ecosystems, and sustainable development frameworks. Temporal mapping demonstrates the field's progression from foundational theoretical debates (2010-2015) to technology-enhanced practical implementations (2016-2020), culminating in current interdisciplinary convergence (2021-2024). Emerging technologies like Industry 4.0 and Artificial Intelligence emerge as critical enablers for digital transformation in collaborative education-industry models. Network analysis reveals strong multidisciplinary intersections, particularly in university-industry-government collaboration mechanisms that drive curriculum innovation and workforce development. Methodological limitations include the WoS database's exclusive coverage, potentially overlooking regionally significant studies from platforms like Scopus or CNKI. While the scientometric approach effectively maps knowledge structures and citation patterns, it lacks a granular examination of implementation challenges – a gap future studies could address through mixed-methods research combining bibliometric analysis with empirical investigations of sector-specific applications. The findings underscore the necessity for expanded data inclusivity and context-specific case analyses to strengthen both theoretical frameworks and practical implementation strategies in this evolving interdisciplinary domain.

REFERENCES

[1] Anaç M, Gumusburun Ayalp G, Karabeyeser Bakan M. A roadmap for reducing construction waste for developing countries[J]. Sustainability, 2024, 16(12): 5057.

[2] Liu X, Liu D, Opoku M. A bibliometric and visualized analysis of meniscus suture based on the WOS core collection from 2010 to 2022: A review[J]. Medicine, 2023, 102(46): e34995.

[3] Gong X. Performance evaluation of industry-education integration in higher education from the perspective of coupling coordination empirical study based on Chongqing [J]. PLOS One, 2024, 19(9): e0308572.

[4] Sun W, Song J, Dong X. Bibliometric and visual analysis of transcranial direct current stimulation in the web of science database from 2000 to 2022 via CiteSpace[J]. Frontiers in Human Neuroscience, 2022, 16: 1049572.

[5] Lu X, Samah N. The evolution and innovation of informational business English education in China: a systematic review between 2019 to 2024[J]. J Hum Centered Technol, 2024, 3(2): 73–83. doi: 10.11113/humentech.v3n2.83.

[6] Zhang M, Chen J. Exploration and research on industry-education integration of vocational education in AI era[J]. Advances in Vocational and Technical Education, 2022, 4(3): 41–47.

[7] Tsang YP, Wu C-H, Ip WH. Exploring the intellectual cores of the blockchain–Internet of Things (BIoT)[J]. Journal of Enterprise Information Management, 2021, 34(5): 1287–1317. doi: 10.1108/JEIM-10-2020-0395.

[8] Wisniewski, D., Zielinski, M., Dabrowski, P. Data-driven approaches to behavior analysis in smart classrooms: A review of AI algorithms[J]. Data Engineering and Applications, 2024, 4(2), 14–30.

[9] Bazel MA, Ahmad M, Mohammed F. Blockchain technology in Malaysian context: bibliometric analysis and systematic review[J]. Int J Adv Sci Eng Inf Technol, 2023, 13(5): 35–67.

[10] Zhao M, Zhang H, Li Z. A bibliometric and visual analysis of nanocomposite hydrogels based on VOSviewer from 2010 to 2022[J]. Frontiers in Bioengineering and Biotechnology, 2022, 10: 914253.

[11] Wang L, Togtokhbuyan L, Yadmaa Z. Visual analysis of the international wellness tourism WOS literature from 1992 to 2019[J]. Int J Spa Wellness, 2021, 4(1): 1–16. doi: 10.1080/24721735.2020.1830527.

[12] He Z, Chen L, Zhu L. A study of Inter-Technology Information Management (ITIM) system for industry-education integration[J]. Heliyon, 2023, 9(9): 1–25.

[13] Masduki NA, Mahfar M, Senin AA. A bibliometric analysis of the graduate employability research trends.[J]. Int J Eval Res Educ, 2022, 11(1): 172–181.

[14] Zhou J, Wang Y, Chen J. Research hotspots and development trends of harvesting robots based on bibliometric analysis and knowledge graphs[J]. Int J Agric Biol Eng, 2024, 17(6): 1–10. doi: 10.25165/ijabe.v17i6.8739.

[15] Wei S, Du T, Zhang L. A comprehensive exploration of astrocytes in migraine: a bibliometric and visual analysis[J]. Eur J Med Res, 2024, 29(1): 321. doi: 10.1186/s40001-024-01919-z.

[16] Huang S, Chen H. Research on quality evaluation of industry-education integration for rural vocational education in the perspective of rural revitalization with the PL-MACONT method[J]. J Intell Fuzzy Syst, 2023, 44(6): 9743–9755.

[17] Brzozka, B. Machine Learning Algorithms in Predicting College Students' Grades: A Review [J]. Journal of Applied Automation Technologies, 2025,3, 1–12.

[18] Ji W, Yu S, Shen Z. Knowledge mapping with CiteSpace, VOSviewer, and SciMAT on intelligent connected vehicles: road safety issue[J]. Sustainability, 2023, 15(15): 12003.

[19] Yan L, Zhiping W. Mapping the literature on academic publishing: A bibliometric analysis on WOS[J]. Sage Open, 2023, 13(1): 21582440231158562. doi: 10.1177/21582440231158562.

[20] Pranckutė R. Web of Science (WoS) and Scopus: The titans of bibliographic information in today's academic world[J]. Publications, 2021, 9(1): 12. doi: 10.3390/publications9010012.

[21] Brabete V, Barbu CM, Cîrciumaru D. Redesign of accounting education to meet the challenges of artificial intelligence–a literature review[J]. Amfiteatru Econ, 2024, 26(65): 275–293.

[22] Zhou X, Jia M, Li W. Higher education or scientific research: Which one contributes more to China's green innovation?[J]. Socio-Economic Planning Sciences, 2024, 94: 101919.

[23] Junjia Y, Alias AH, Haron NA. A bibliometric review on safety risk assessment of construction based on CiteSpace software and WoS database[J]. Sustainability, 2023, 15(15): 11803.

[24] Yonfá-Medranda M, Sabando-Vera D, Pérez-Gálvez J. Consumer ethnocentrism: bibliometric analysis and literature review through cognitive structure and mapping of research based on the web of science (WoS) and scopus[J]. Cogent Business & Management, 2024, 11(1): 2377320. doi: 10.1080/23311975.2024.2377320.

[25] Zuo Z, Cheng J, Guo H. Knowledge mapping of research on strategic mineral resource security: A visual analysis using CiteSpace[J]. Resour Policy, 2021, 74: 102372.

[26] Laura Icela G-P, María Soledad R-M, Juan Antonio E-G. Education 4.0 maturity models for society 5.0: Systematic literature review[J]. Cogent Business & Management, 2023, 10(3): 2256095. doi: 10.1080/23311975.2023.2256095.

[27] Yuan P, Yang X. Exploration of the model of deepened industry–education integration in the digital economy era[J]. Journal of Internet and Digital Economics, 2024, 4(3): 179–186.

[28] Rakhimovna BN. Integration of theory and practice of the dual education system in the field of light industry education[J]. Eur Int J Multidiscip Res Manag Stud, 2024, 4(2): 336–341.

[29] Chen Y-H, Yin M-Q, Fan L-H. Bibliometric analysis of traditional Chinese medicine research on heart failure in the 21st century based on the WOS database[J]. Heliyon, 2023, 9(1). doi: 10.1016/j.heliyon.2022.e12770.

[30] Sarkar A, Wang H, Rahman A. A bibliometric analysis of sustainable agriculture: based on the Web of Science (WOS) platform[J]. Environ Sci Pollut Res, 2022, 29(26): 38928–38949. doi: 10.1007/s11356-022-19632-x.

[31] Li H, Li B. The state of metaverse research: a bibliometric visual analysis based on CiteSpace[J]. J Big Data, 2024, 11(1): 14. doi: 10.1186/s40537-024-00877-x.

[32] Chen Z, Zhang W, Li L. Evaluation of urban industry-education integration based on improved fuzzy linguistic approach[J]. Huang G. ed. Mathematical Problems in Engineering, 2021, 2021: 1–14. doi: 10.1155/2021/6610367.

[33] Chi Z, Liu Z, Wang F. Driving circular economy through digital technologies: Current research status and future directions[J]. Sustainability, 2023, 15(24): 16608.

[34] Li J, Xue E. Investigating international students' cultivation system for higher education sustainability in China: stakeholders' perspectives[J]. Sustainability, 2022, 14(22): 15335.

[35] Ding X, Yang Z. Knowledge mapping of platform research: a visual analysis using VOSviewer and CiteSpace[J]. Electron Commer Res, 2022, 22(3): 787–809. doi: 10.1007/s10660-020-09410-7.

[36] Fang X, Zhang Z. Hotspots and trends in health-oriented physical literacy research: A visual analysis based on the WOS database[J]. BMC Public Health, 2024, 24(1): 1480. doi: 10.1186/s12889-024-18951-7.

[37] Feng X, Wang X, Qi M. A comparison study on innovation and entrepreneurship education in and out of China from a bibliometric perspective[J]. Libr Hi Tech, 2024, 42(6): 1810–1838.

[38] Li J, Ma W, Dai X. China's policy environment's development and path from the perspective of policy sustainability: A visual analysis based on CNKI and WoS[J]. Sustainability, 2022, 14(24): 16435.

[39] Shi R, Wan X. A bibliometric analysis of knowledge mapping in Chinese education digitalization research from 2012 to 2022[J]. Humanit Soc Sci Commun, 2024, 11(1): 1–14.

[40] Gao Y, Lin R, Lu Y. A visualized analysis of the research current hotspots and trends on innovation chain based on the knowledge map[J]. Sustainability, 2022, 14(3): 1708.

An Efficient and Scalable Reinforcement Learning-Driven Intelligent Resource Management and Secure Framework for LoRaWAN

Shaista Tarannum[1], Usha S.M[2]
Department of Electronics and Communication Engineering, JSS Academy of Technical Education, Bengaluru, India[1]
Visvesvaraya Technological University, Belagavi-590018, Karnataka, India[2]

***Abstract*—This study proposes a Q-learning-based adaptive duty cycle scheduling algorithm for LoRaWAN in a smart city eco-system to enhance the energy efficiency, reduce transmission delay, and handle dynamic traffic conditions. Additionally, it also incorporates an intelligent and efficient channel utilization scheme for LoRaWAN-enabled IoT networks and also integrates a lightweight security strategy at the edge (gateways), making it suitable for low-power, low-computation LoRaWAN environments. In this adaptive and intelligent LoRaWAN framework Q-learning agent dynamically selects various transmission actions based on the contextual states, including buffer size, energy levels, and channel conditions, which optimizes energy efficiency and also enhances the reliability of data transmission in LoRaWAN. The light-weight intrusion detection mechanism also filters suspicious packets using trust scores and payload analysis to ensure secure data delivery and adaptive, scalable, and proactive protection against several prevalent threats in LoRaWAN-driven IoT. It also incorporates a channel-aware scheduling to avoid congestion and improve overall transmission performance. Experimental outcome further confirms improvement over throughput, delay, bandwidth utilization, energy conservation, and resilience against malicious or faulty transmissions, demonstrating the framework's ability to optimize the resource allocation performance while balancing the above metrics adaptively.**

***Keywords*—LoRa; LoRaWAN; Q-learning; adaptive duty cycle; channel scheduling; energy efficiency; intrusion detection; trust score; resource management; IoT security**

I. Introduction

The rapid advancement of the Internet of Things (IoT) has created a demand for low-power, long-range wireless communication technologies. IoT involves connecting a wide range of end-devices, which are battery-powered sensor nodes, to the Internet and also enabling communication and data exchange between them [1] [2]. Therefore, the need arises to design the power usage profile carefully in order to extend the battery's lifetime. Also, the communication range needs to go from several hundred meters up to several kilometers as IoT end devices are distributed over a large area of operation. Considering all the aforementioned characteristics, this can be only realized by using the low power wide area network (LPWAN) technologies as LPWAN technologies can support resource management, throughput, and delay constraints in IoT [3]. There exist several LPWAN technologies present in the market, such as SigFox [4], Narrow Band (NB)-IoT [5], or Long-Range Wide Area Networks (LoRaWAN) [6] and Long-Term Evolution for Machines (LTE-M) [7]. The maximum data rate in SigFox is ~100 bps, whereas in NB-IoT it is ~250 kbps. On the other hand, in LoRaWAN, it is approximately ~50 kbps, and in the case of LTE-M, the maximum data rate is approximately ~1 Mbps. LoRaWAN [6] is among the leading LPWAN technologies as it offers the possibility for private network deployments and easy integration with a number of worldwide network platforms and has the ability to provide long-range communication with low power consumption. Due to this and its open access specification, LoRaWAN have gained significant attention from academia and industries for IoT [8] [9]. The LoRa physical layer has been patented by Semtech in the year 2014 [10]. LoRa is a radio frequency (RF) modulation technology that defines the physical layer features for long-range communications. However, the LoRaWAN medium access control (MAC) protocol is an open-source protocol standardized by the LoRa Alliance that runs on the top of LoRa physical layer.

LoRaWAN is also popular in smart city applications as it provides cost-effective, scalable, and power-efficient solutions for connecting thousands of widely distributed and low-data-rate devices. It supports long-range communication while covering ~2-15 km in rural areas and 1-5 km in urban settings. The end devices in LoRaWAN can last upto 5 to 10 years on battery, that reduces the need for frequent maintenance. It also offers low-cost unlicensed spectrum (ISM) bands, whereas LoRaWAN gateways are also affordable and can support thousands of devices. It also supports municipalities to deploy and control their own infrastructures in smart city applications. LoRaWAN also supports flexible network architecture with public, private, and hybrid networks, whereas designed for massive IoT deployments. Also, it uses AES-128 encryption schemes at the network and application layers while ensuring end-to-end security for data transmitted from sensors to applications [9] [10]. LoRaWAN is a popular communication protocol designed for low-power wide-area IoT deployments. While it excels in range and energy efficiency for small data transfers but it faces several inherent limitations and challenges due to its design trade-offs [20] [21]. The challenges arise due to low data rates, inefficient resource allocation, energy consumption, and network congestion problems [22]. It has also been observed that traditional fixed-duty cycle mechanisms suffer from suboptimal bandwidth utilization, increased transmission delays, and energy wastage, especially

in dynamic environments where real-time data transmission is crucial [23]. The root cause for the low data rate transmissions is chrip spread spectrum modulation, which prioritizes robustness and range over speed, and also spreading factor (SF) that results in long transmission time and poor latency [24]. LoRaWAN uses a pure ALOHA channel access method that also increases the chances of retransmissions. collisions and unfair resource usage. The frequent transmissions (TX), collisions, and high SF factors result in reduced battery life. LoRaWAN gateways operate in licensed ISM bands (868 MHz in EU, 915 MHz in US) with strict duty cycle limitations. As the network of IoT end devices grows with limited channel availability, that also results in frequent packet drops, low QoS, and scalability problems [25]. It is also observed that the LoRaWAN protocol suffers from security vulnerabilities to jamming and intrusions. No such in-built mechanism is found in LoRaWAN for intrusion detection or anomaly classification [15].

Machine Learning (ML) is a popularly growing field with many applications, including wireless communications [26] [27]. There are various studies which claim that ML could be used to improve the performance, efficiency, and security of LoRaWAN [28]. ML algorithms empower LoRaWAN networks to dynamically allocate resources, predict network traffic, mitigate interference, and optimize energy consumption, thereby enhancing network capacity, reliability, and battery life. With ML-driven insights, operators can proactively plan network expansions and ensure better quality of service (QoS), and also achieve self-optimizing networks that autonomously adapt to changing conditions. However, popular supervised learning approaches such as SVM, Decision Trees, Random Forest, and Logistic Regression models rely on a large amount of labelled data, which is often scarce and expensive in LoRaWAN due to limited sensing and reporting capability [29]. Also, models trained on static datasets may fail to generalize in dynamic, noisy environments with high traffic variability. Traditional supervised approaches also do not adapt well to real-time changes in topology or energy levels. These models also suffer from scalability issues and security bias limitations. It has been also observed that existing Deep Learning (DL) models, such as CNN, LSTM, DNN, auto encoders, require significant resources for training and interference which is not practical for low-powered LoRa nodes. DL models heavily rely on large datasets which is often unavailable in LoRaWAN due to low data rates and sparse feedback. Also, these DL models suffer from latency issues, overfitting problems, black-box nature, and deployment complexities [29] [30]. It has been also observed that the existing unsupervised ML techniques suffer from poor contextual awareness, false positives, and parameter tuning issues in large-scale LoRaWAN deployments.

The proposed system, therefore, aims to optimize the throughput, minimize transmission delay, and preserve energy while meeting the criteria for secure transmission, thereby enhancing the overall performance of LoRaWAN. However, the proposed work also finds the popularity of RL strategies and their scope towards improving the performance of LoRaWAN and further formulates a unique form of light-weight computational and analytical framework for Q-learning based adaptive duty cycle management in LoRaWAN, which is also further integrated with a security modeling. Here, the framework aims to optimize the LoRaWAN resources in dynamic conditions and also offers resiliency against different forms of adversaries in LoRaWAN, which affects the energy and QoS performance in LoRaWAN.

The key contribution of this work is listed as follows:

- Unlike traditional fixed duty cycling and static scheduling strategies, the proposed work employs Q-learning to dynamically adapt transmission decisions based on real-time network conditions and also significantly improves energy efficiency and throughput. Here Q-learning based adaptive duty-cycling dynamically adjusts transmission slots considering buffer, energy, channel, and duty cycle constraints.
- In contrast to the generic ML models that require extensive training data and centralized models, the proposed approach offers a light-weight and model-free reinforcement learning strategy that enables distributed and online learning at individual IoT nodes in LoRaWAN.
- The integration of trust-based intrusion detection modeling enhances the security by proactively filtering malicious data packets, a feature that is missing in existing standard ML approaches for resource management in LoRaWAN that focus solely on performance optimization without considering security. The framework also can be extended to incorporate trust scores from gateway-based intrusion detection, allowing the RL agent to make security-aware scheduling decision without compromising resource optimization.
- It offers an effective RL-driven approach for selecting transmission strategies that adaptively optimizes throughput, bandwidth utilization in response to dynamic traffic patterns. The proposed work shows superior performance in terms of delay, energy conservation, and secure transmission under variable traffic and channel conditions compared to popular baseline schemes.
- Unlike existing RL-based LoRaWAN studies, our framework uniquely integrates a Q-learning-driven adaptive duty cycle with trust-aware security with trust-aware security feedback. It combines Q-learning driven adaptive duty cycle for efficient resource management with a trust-aware security module. While the Q-learning agent optimizes transmission slots and power levels, the security layer independently evaluates trust scores from gateway-based intrusion detection that ensure reliable and secure data transmission alongside optimized network performance.
- In future work, the Q-learning agent could be extended to incorporate trust scores from gateway-based intrusion detection, which will enable joint resource management and security-aware decision making.

The aforementioned contributions are presented in a structured manner. The organization of the study is as follows: Section II reviews the existing methodologies, while the identified limitations are outlined in Section III. Section IV describes the system model, followed by the discussion of results in Section V. Finally, Section VI concludes the study with a summary of key findings.

II. Review of Literature

There exist various forms of related research studies that have also focused on improving the performance of LoRaWAN using ML approaches. The authors in [11] have proposed a load-balancing method for dense IoT networks such as smart city scenarios. The authors basically performed training of various ML techniques such as Multiple Linear Regression (MLR), Gaussian Naïve Bayes (GNB), Linear Discriminant Analysis (LDA), Decision Tree (DT), Random Forest (RF), and few others. These classifiers are applied to an urban IoT network, where the simulation results showed performance improvement over packet success ratio (PSR), and the framework also offered optimized energy consumption in the LoRaWAN network. In [12], the authors have introduced an SF allocation scheme considering SVM and DT that aims to resolve the collision issue in the LoRaWAN network. The training dataset was generated using simulator for LoRa SimLoRaSF, and a custom simulator was also designed for LoRaWAN using Python. The authors in [13] also emphasizes toward resource classification problem for static EDs in LoRaWAN using various ML techniques. Such as RF, SVM, logistic regression (LR), K-Nearest Neighbour (KNN), and few more others. The experimental outcome shows that the RF method accomplished the highest accuracy of 92% compared to other ML techniques. It has been observed that various DL methods were applied to improve the performance of LoRaWAN by optimizing resource parameters, predicting network traffic, mitigating inter-intra interferences, and optimizing energy constraints.

An extended Kalman Filter-based LSTM method based on regression is proposed for predicting collisions in the LoRaWAN network, which is introduced in the study of [14]. As a collision in the LoRa network is directly linked with the SF, hence SF has not been considered for adaptive configuration. As a result, the presented LSTM approach leads to underperformance when applied in a dynamic LoRaWAN network. The study in [15] proposed DeepLoRa, which is an environment-aware path loss model. On the other hand, in [16], a DL method is proposed for managing the transmission interval of IoT devices in LoRa networks by utilizing Intel Berkeley Research Lab Data. The paper [17] proposed a DL method for joint collision detection and resource management. Here, the presented work considers two DL methods: fully connected neural networks (FCNNs) for collision detection and CNN for SF management. The experimental results show improved prediction performance and energy consumption when compared with traditional ML methods. The paper [18] proposed an RL-based approach for optimizing and updating LoRa communication parameters. The authors utilized the RL method to derive optimal disseminating policies by aiming to maximize the accumulated average node throughput. The authors claim that their approach to the LoRaWAN has given a remarkable increase in the accumulated average per-node throughput of 147%.

The authors in [19] presented a novel method for resource allocation in LoRaWAN networks. Here, the method used Q-learning strategy of RL to learn the optimal resource allocation policy for each ED in the network. In the presented method, GW acts as an agent of Q-learning, where the Q-reward is based on the weighted sum of the number of successfully received packets in the proposed method. The Q-learning method was evaluated using simulation, and the outcome shows that the proposed method achieves PSR by ~20% compared to a random resource allocation scheme.

III. Research Problem

LoRaWAN networks significantly face challenges in balancing energy efficiency, reliable data transmission, and security against malicious activities. Traditional static duty cycle mechanism in LoRaWAN leads to inefficient bandwidth utilization and energy wastage, especially under dynamic traffic and channel conditions. Furthermore, LoRaWAN's lightweight nature makes it susceptible to various security threats, including spoofing and packet injection. The research addresses the need for an adaptive secure scheduling mechanism by formulating a Q-learning-based framework that can dynamically adjust the transmission duty cycle based on real-time state conditions. However, the challenge also arises to design a light-weight trust-aware intrusion detection system to block malicious packets in LoRaWAN. Despite the popularity, existing AI/ML-based optimization techniques often fail to adapt to the variability in network traffic, dynamic occupancy patterns, and energy constraints in large-scale LoRaWAN deployments.

IV. Research Method

The proposed work, therefore, aims to develop a novel, highly efficient, and scalable AI/ML-driven adaptive resource management and secure framework for LoRaWAN in a smart city eco-system where the resource management strategy is well-capable of optimizing the duty cycle allocation, bandwidth utilization, and energy-efficient transmission while maintaining high data throughput and minimal latency. The proposed framework of LoRaWAN operates in the MAC layer that effectively defines how LoRa devices communicate with the gateway and network servers. In the proposed work, the adaptive LoRaWAN framework (Fig. 1) is structured into three primary components, which are Q-learning-based scheduling, channel-aware management, and lightweight edge security. Here, Fig. 1 also shows a typical LoRa network deployment scenario. Before delving into algorithmic details, the section outlines the operational flow of these core components in LoRaWAN.

A. LoRaWAN-Based IoT Deployment

Let, $D = \{d_1, d_2, \ldots, d_n\} \in \mathbb{Z}^+$ be the set of n LoRa IoT devices in a LoRaWAN environment. The LoRa-based IoT network also consists of m LoRa gateways such as $G = \{g_1, g_2, \ldots, g_m\} \in \mathbb{Z}^+$. It is also assumed during the analytical modeling that the network should also consist of a central network server S. The total time steps for simulation is considered to be T. Therefore, each IoT device $d_i \in D$ is

located at position of $\mathrm{Loc}(d_i) \in \mathbb{R}^2$. And also, each gateway $g_j \in G$ has a coverage radius of R_j. The deployment assumption also considers that the devices and gateways are randomly distributed in a 2D space, and coverage check can be validated considering the following Eq. (1).

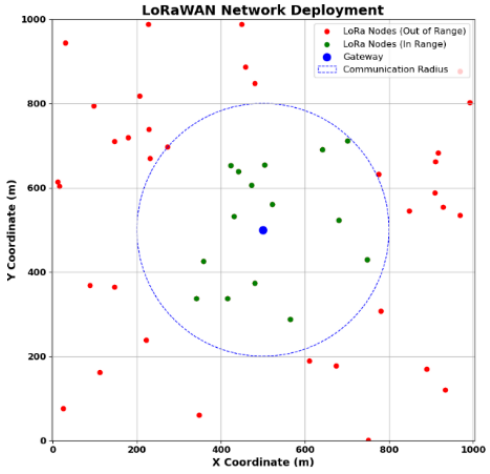
$$\left\|\mathrm{Loc}(d_i) - \mathrm{Loc}(g_j\right\| \leq R_j \quad (1)$$

At each time step of $t_i \in T$ each device d_i generates a data packet in the form of $P_i^{(t)} = \langle temp_i^{(t)}, hum_i^{(t)}, \tau_i^{(t)}, id_i \rangle$ where $temp_i^{(t)}$ represents the temperature readings, $hum_i^{(t)}$ represents the humidity readings, $\tau_i^{(t)}$ represents the current time stamp and id_i refers to the unique device ID. It also assumes that the packet $P_i^{(t)}$ is sent to a randomly selected gateway $g_j \in G$. Here it is assumed that IoT end devices periodically generate data packets, and that are transmitted to randomly selected gateways. The packet loss model also considers that each gateway g_j has a probability of packet loss of 0.05 and success of 0.95. The condition of data transmission is modeled using Eq. (2). If it is found that $X_{ij}^{(t)} = 1$, then the gateway forwards the packet to the server $P_i^{(t)} \rightarrow S$. The server further stores and processes the packets in Eq. (3).

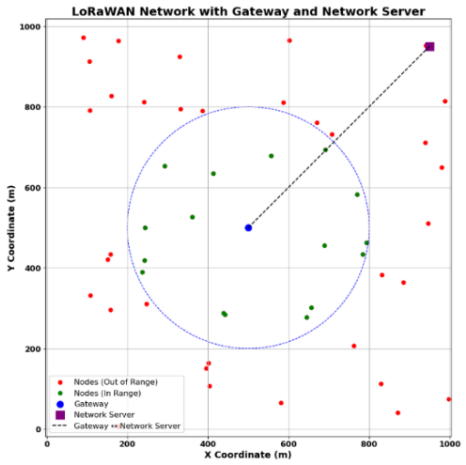
$$X_{ij}^{(t)} = \begin{cases} 1, & \text{if } g_j \text{ succesfully receives } P_i^{(t)} \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

$$\xi_s = \cup_{t=1}^{T} \left\{ P_i^{(t)} | X_{ij}^{(t)} = 1 \right\} \quad (3)$$

Here, in the above Eq. (3), ξ_s represents the cumulative set of successfully received packets over time from all IoT end devices in the LoRaWAN-enabled smart city eco-system. The proposed work emphasizes towards designing and developing a novel adaptive duty cycle slot allocation mechanism using Reinforcement Learning (RL) to enhance the performance of large-scale LoRaWAN in smart city eco-system. The following Fig. 1(a) shows the LoRaWAN network deployment scenario without a network server, and Fig. 1(b) shows the LoRaWAN network deployment scenario with the inclusion of a network server in the proposed work.



(a)



(b)

Fig. 1. LoRa network deployment in IoT smart city: a) LoRaWAN network with gateway placement, b) LoRaWAN with gateway and network server.

B. Improving LoRaWAN Performance Using Reinforcement Learning (RL) Algorithms

The proposed study emphasizes towards improving the performance of LoRaWAN via ML-based decision making for adaptive duty cycle slot allocation. The proposed work leverages RL techniques in LoRaWAN to intelligently manage the network resources and aims to maximize efficiency along with optimal delivery performance. Reinforcement Learning (RL) is a data-driven approach where it learns rules and policies from experience by interacting with the network and observing the results. RL is a dynamic approach that has the capability to adapt to the environmental changes. LoRaWAN networks are constantly changing owing to the dynamic scenarios and due to the underlying propagation environment. As a result, RL algorithms can be used to learn how to optimize the network parameters for these changes, ensuring that the network remains reliable and efficient. As RL is a scalable approach, it can be used to optimize large and complex networks [21]. Therefore, the proposed work realizes that in LoRaWAN, RL could be used to optimize resources by training agents for making decisions to maximize overall network efficiency and minimize interference, along with energy consumption. The proposed work in the first phase of design, therefore, introduces a novel and cost-effective adaptive duty cycle slot allocation using Q-learning modeling and also ensures dynamic slot assignment. In the second phase of design, it incorporates a scheme for bandwidth and efficiency optimization, followed by a security analysis in the third phase.

1) Adaptive duty cycle with RL in LoRaWAN: Duty cycle in LoRaWAN determines how often a device is capable of transmitting data packets. It has been observed that the static allocation paradigms often lead to congestion, collisions, or under-utilization of resources. Therefore, the proposed work introduces an approach of adaptive duty cycling using Q-learning to dynamically choose the best time slot based on traffic, battery, and buffer status, which in longer run reduces contention and also improves energy efficiency.

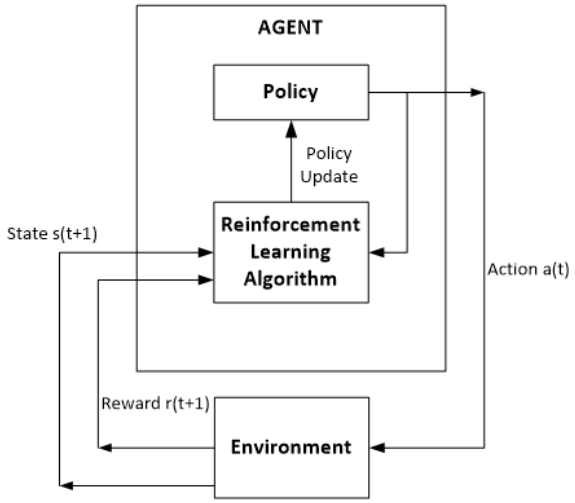


Fig. 2. Reinforcement learning in LoRaWAN.

2) State representation for Q-learning agent: In the proposed analytical modeling of LoRaWAN, each end device state in the context of IoT smart city is defined as a tuple of $s_t = \langle b_t, E_t, l_t, D_t, c_t \rangle$. Here, b_t refers to buffer size at time t, E_t implies the battery level at time t, whereas l_t implies time since last transmission. D_t on the other hand, implies current duty cycle slot and c_t represents channel status (i.e 0=idle, 1 = busy). Fig. 2 shows baseline RL concept which is adopted in the proposed LoRaWAN framework designing.

3) Action space definition in LoRaWAN: The proposed analytical strategy further defines the action space $A = \{TX_{high}, TX_{low}, wait\}$ as a set of high power transmission, which require high energy usage and lower delay denoted with TX_{high}, low power transmission, which require low energy usage but higher delay denoted with TX_{low}, and $wait$ implies skip transmission where the buffer grows and battery is conserved. The prime reason behind considering TX_{high} is that it is used for urgent data or poor channel conditions. It aims to minimize the delay and improve packet delivery probability (stronger signal) for urgent packets or unreliable channel conditions in LoaRaWAN. On the other hand, the action corresponds to TX_{low} imply when channel conditions are good or delay is tolerable for packet transmission. This action also conserves battery life while still sending data. Here, the action $wait$ refers to preserve battery or adhere to duty cycle limits. This action has got importance when the network is highly congested or data is not urgent. This approach not only conserve energy but also helps avoiding duty cycle violation. In the RL perspective the proposed work decides the action space to support light-weight decision model in resource constrained devices with three discrete actions. The learning agent learns a policy to select the best action at each time step depending on the above state space criteria highlighted in $s_t = \langle b_t, E_t, l_t, D_t, c_t \rangle$ to enhance the performance and resource management in LoRaWAN. In the proposed work Q-learning helps managing the resources and LoRaWAN traffic and enables adaptive traffic handling.

4) Q-learning in adaptive traffic handling: In the proposed approach of performance improvement of LoRaWAN the states encode the traffic contexts in the form of s_t. Here, the buffer size implies traffic backlog, which means how much data is waiting to be sent. On the other hand, E_t refers to the remaining battery level of end IoT devices. On the other hand, l_t helps controlling the latency factor. The factor of D_t is used as a duty cycle slot that measures the compliance constraint. These states help the learning agent understand the traffic load and network status accordingly RL mechanism effectively manages the resources. TX_{high} is learned when urgent traffic and high buffer build-up occur. On the other hand, TX_{low} is learned by the agent when traffic is moderate and conditions are favourable. The Q-learning learns to skip Tx for the action $wait$ and avoid penalty and transmit later when the condition improves. In the case of learning$wait$, the Q-learning model realizes that the traffic is low, that means there is no urgency to transmit, whereas transmission at that point might waste energy or duty cycle. $wait$ allows agents to accumulate more data and possibly send it together later. Q-learning learns that delaying when the buffer is small and doesn't incur a big penalty, and also may result in higher future rewards. The $wait$ is useful as it helps avoiding duty cycle violation, helps when the channel is busy ($c_t = 1$) and also useful in the case of battery E_t is critically low. The agent learns to wait for better conditions. Here, via Q-learning, the agent chooses the optimal actions to balance delay, energy usage, and duty cycle compliance, which also positively influence the throughput and data delivery outcome. The Q-Function updating is represented with Eq. (4).

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha \left[r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right] \quad (4)$$

Here, α represents the learning rate, whereas γ implies the discount factor, and also r_t refers to the intermediate rewards at time t. Also, in the above expression, a' represents the next action, and further it leads to the Q-value for state action pair in $Q(s_t, a_t)$. While iteratively updating the above expression (4), the agent progressively learns when to transmit and at what power in LoRaWAN, and also learns when to wait to avoid congestion or conserve resources. It also adapts to dynamic traffic patterns in real-time. The agent evaluates b_t and if it is high, then it immediately sends in TX_{high} or TX_{low} depending upon the channel conditions. If E_t is low, then the agent prefers to conserve energy, so either it chooses or TX_{low} or $wait$. If l_t is high, then it avoids excessive delay and prefer any Tx. If D_t is found near max, then it avoids sending and prefers $wait$. Also, if $c_t = 1$ that implies the channel is busy and prefers $wait$. Here, $D_t \sim 1$ indicates that the end IoT node has used almost all of its allowed transmission time for the current regulatory duty cycle window. That's why it is safer to avoid transmitting until the cycle resets. In LoRaWAN, especially in the EU 868 MHz band is subject to strict regulatory duty cycle limits. If $D_t \sim 1$ in max and the learning agent still chooses to transmit that means it may get penalty, drop the packet and delay can occur for future transmissions. This results in low or negative reward, so over time the agent

learns that $wait$ is the optimal action, when $D_t \sim 1$. With approach the Q-table find which action yields the best long-term return and balances throughput, energy and reliability.

5) Reward function modeling: The design of reward function modeling encourages high rewards for timely transmission with low power. And also penalizes the agent for buffer overflows, delays, or duty violations. It is modeled using Eq. (5). Here, e = 5 for TX_{high} and e = 2 for TX_{low}:

$$r_t = \begin{cases} 10 - e & If\ TX\ succesful \\ -0.5 & if\ wait \\ -2 & if\ TX\ failed\ or\ battery\ low \end{cases} \quad (5)$$

The reward modeling in LoRaWAN encourages the learning agent with a high reward if fast delivery is needed. On the other hand, TX_{low} also gets a higher reward despite a higher delay, which promotes energy efficiency. Whereas TX_{high} is used when speed is critical but very less frequently. $wait$ is not rewarded, but the penalty is low, allowing it to be chosen in risky or sub-optimal states. The proposed Q-learning in LoRaWAN also penalizes risky transmissions, such as if Tx is fails or Tx when the battery is low. In such cases, the agent gets negative rewards, and further, this reward structure also pushes the agent to avoid such states.

6) LoRaWAN adaptive duty cycle mechanism: In the proposed work, the time is divided into 10 discrete duty cycle slots in the form of D_t. Here, each slot might represent a fixed duration, and the model cycles through them as time advances. The update rule for adaptive duty cycle is given using Eq. (6):

$$D_{t+1} = D_t + 1\ mod\ 10 \quad (6)$$

This equation shows that after each decision step (time step), the slot automatically advances to the next. In the proposed system content of LoRaWAN, the model behaviour simulates a rolling time window where duty cycle behaviour resets periodically, matching regulatory constraints (e.g. 1% per hour transmission time in EU868). In the proposed Q-learning-based adaptive duty cycling-based strategy, agents learn which slots are best for transmission based on success rate (i.e., whether Tx is successful), energy cost (TX_{high}, TX_{low}) , and channel conditions (e.g., fewer collisions). Through Q-learning, LoRaWAN builds state-action values $Q(s_t, a_t)$ that include D_t as a state component. It has to be noted that with this approach, the agent doesn't learn what to do, but it learns when to do it. The benefit of this adaptive duty cycling approach is that the agent learns to adapt to traffic patterns, interference, and energy constraints over time.

It has to be noted that the proposed framework operates in three phases: 1) Q-learning-based adaptive duty cycle assignment, 2) efficient channel scheduling, and 3) security-aware packet evaluation. Algorithm 1 to Algorithm 3 together represent the complete operational procedure in which the RL agent selects optimal transmission actions while considering energy levels, buffer status, and channel conditions, thereby achieving adaptive, secure, and efficient resource management in LoRaWAN. The RL agent is also capable of selecting optimal transmission actions while considering trust scores as well. The following algorithm shows Q-learning for adaptive duty cycle in LoRaWAN.

Algorithm 1: Q-learning for Adaptive Duty Cycle in LoRaWAN

Initialize $Q(s_t, a_t)$ arbitrarily for $\forall \langle s_t, a_t \rangle$ pairs
For each episode:
  Initialize environment (E) and state $s_t = \langle b_t, E_t, l_t, D_t, c_t \rangle$
  Repeat for each time step
    With probability ε select random action
    $a_t \in \{TX_{high}, TX_{low}, wait\}$
    Otherwise select $a_t = \arg\max_{a'} Q(s, a)$
    Execute action a, observe
      reward r_t,
      next state s_{t+1}.
      $D_{t+1} = D_t + 1 \bmod 10$ ← Adaptive Duty Cycle slot update
    Update Q(s,a) using:
      $Q(s_t, a_t) = Q(s_t, a_t) + \alpha \left[r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right]$
      $s \rightarrow s'$
      If terminal condition is met, break

The algorithm design and modeling ensure that the Q-learning agent adapts based on the dynamic state s_t. It also offers a duty cycle slot allocation where the reward strategy is encouraged for optimal time. Here, Q-learning automatically adapts policy using reward signals.

7) Optimization of bandwidth and data efficiency in LoRaWAN: It has to be noted that LoRaWAN in smart city eco-system suffers from collisions and idle listening as many IoT end devices share the same channel. Therefore, optimizing channel utilization (bandwidth) and efficient data scheduling can significantly improve the performance of LoRaWAN. If there are N number of end devices (ED) and P_t indicates the packets the successfully transmitted packets at time t ,whereas $\mathcal{U}_t$ represents the bandwidth utilization at time t. Then the bandwidth utilization can be modelled using Eq. (7) and Throughput T_t is measured using Eq. (8). The proposed work further evaluated the transmission delay using Eq. (9). Here, P_t represents the set of packets receiver at t and t_{rx}^p, t_{gen}^p represents the generation time and reception time of packets. The system also computes the battery energy level of IoT nodes or end devices in the form of E_i^t. Here, E_i^t represents the battery level of device i at step t.

$$\mathcal{U}_t = {}^{P_t}/_{N} \quad (7)$$

$$T_t = \sum_{i=1}^{N} \delta_i^t \text{ where } \delta_i^t = \begin{cases} 1 & if\ node\ i\ transsmission\ was\ succesful\ at\ t \\ 0 & otherwie \end{cases} \quad (8)$$

$$Delay_t = {}^{1}/_{|P_t|} \sum_{p \in P_t} (t_{rx}^p - t_{gen}^p) \quad (9)$$

In the proposed LoRaWAN framework Q-learning chooses TX_{high} and TX_{low} based on the congestion and energy. The bandwidth utilization is also tracked via $\mathcal{U}_t$ whereas the

proposed reward structure and simulation loop encourages maximized throughput. The proposed Algorithm 2 for efficient channel scheduling is represented as follows:

Algorithm 2: Efficient Channel Scheduling in LoRaWAN

For each time slot t:
  For each IoT node i:
    Estimate channel status c_t
    If $b_t > \phi$ and energy sufficient:
      Select TX_{high} and TX_{low} via Q-learning
    Else:
      Choose the $'wait'$ action
    Update packet success/failure
    Update metrics:
      Throughput, delay, bandwidth utilization

It has to be noted that the proposed work in LoRaWAN approximates an efficient channel scheduling (ECS) strategy by using a Q-learning agent to dynamically decide on transmission power or WAIT based on local buffer status, energy, duty cycle, and channel status. However, full-fledged ECS features like multi-channel assignment, real-time congestion estimation are marked as future extensions.

C. Security-Aware RL Approach in LoRaWAN

The extensive analysis on LoRaWAN shows that it is vulnerable to different forms of attacks, such as spoofing, flooding, and data injection. Here, in the proposed system, a security-aware RL strategy can detect malicious data patterns, it can also maintain trust scores, and drop packets based on suspicious behaviour. It also ensures secure routing and packet forwarding. The security modeling considers a functional strategy of malicious packet indicator $\phi(P_i^{(t)})$ which generates binary flags if packet $P_i \in s$ is spoofed or malicious, else it flags to 0. It can be represented using Eq. (10). If s_t denotes the set of received packets at time t. Then the intrusion detection rate is computed using Eq. (11):

$$\phi(P_i^{(t)}) = \begin{cases} 1 & if\ P_i^{(t)}\ is\ malicious \\ 0 & otherwise \end{cases} \quad (10)$$

$$I_t = {}^{1}\!/_{s_t} \sum_{P_i^{(t)} \in s_t} \phi(P_i^{(t)}) \quad (11)$$

The intrusion detection rate basically measures how many suspicious packets were found in the current time window. The value of I_t implies that the network is under attack or faulty devices are sensing data. In this proposed security-aware reinforcement learning approach, each IoT device is continuously monitored. And also, the system learns patterns of normal versus suspicious behavior. The devices that behave badly get low trust scores, and their packets are blocked. Here, the system scans the packets from all the IoT devices, and it incorporates a model to spot unusual behavior in a packet, such strange timing or values. The proposed work also implements a security filter in this adaptive resource management framework of LoRaWAN. Here, each gateway uses a rule-based intrusion detection function using Eq. (12). Here, T_d indicates the temperature reading, H_d indicates the humidity reading, whereas T_{max} = 100 degrees and H_{min} is ~5%. The trust score τ_i for IoT device i is evaluated using Eq. (13). The τ_i ranges between 0 and 1. Here 1 means fully trusted device, whereas 0 indicate highly suspicious device. When too many of a device's packets are flagged, then the trust score drops. The packet drop rate δ-drop is also estimated using Eq. (14). This mathematical model evaluates how many packets are rejected as the proposed framework could not trust the source device.

$$Detect(d) = \begin{cases} 1 & \text{if } T_d > T_{max}\ or\ H_d < H_{min} \\ 0 & otherwise \end{cases} \quad (12)$$

$$\tau_i = 1 - {}^{\text{Flagged } P_i^{(t)} \text{ from i}}\!/_{\text{Total } P_i^{(t)} \text{ from i}} \quad (13)$$

$$\delta_{drop} = {}^{P_i^{(t)} \text{ dropped due to low } \tau_i}\!/_{\text{Total received } P_i^{(t)}} \quad (14)$$

In the proposed work and security strategy, the model simulates malicious behavior in LoRaWAN while injecting fake data. It also simulates malicious IoT devices sending falsified or abnormal data, which represents active attacks like data injection. The proposed security filtering and black listing formulation is well capable of defending network flooding attacks or data corruption with fake sensor values. It also prevents the adversaries from resending previously valid packets. As the gateway/server could be extended to discard old or replayed packets by checking if the timestamp is within a valid range or not. If a node continuously refuses to transmit data, then the proposed security framework could be extended with a Q-learning agent that discourages excessive wait or non-transmitting states via negative rewards and mitigates routing blackhole by adapting duty cycles and penalizing inactivity in LoRaWAN. The system is also well capable of resisting Grayhole attack, Jamming attack, Sybil, and DoS intrusions in LoRaWAN. The next segment of the study further illustrates the results and discussion on the outcome obtained from simulating the above analytical algorithms through numerical computing and analysis.

Algorithm 3: Intrusion Detection and Secure Transmission in LoRaWAN

For each received $P_i^{(t)}$ from device i
  1. Extract key info: device_id, timestamp, data content (payload)
  2. Run anomaly or unusual behaviour detector on $P_i^{(t)}$
    - If $P_i^{(t)}$ is suspicious or trust_score $[i] < 0.6$
      Drop the packet
      Reduce trust_score$[i]$ by a penalty value.
      Blacklist ← Blacklist ∪ {Device$_i$}
    - Else
      Accept the $P_i^{(t)}$
      Increase trust_score $[i]$ by a reward value.

It has to be noted that in the proposed framework the security layer operates synergistically with the Q-learning agent, in which the trust score of each device is periodically updated considering packet consistency and transmission success rate. These trust updates in longer run will indirectly influence the agent's state vector and also enable adaptive scheduling decisions that prioritize reliable and trusted nodes within LoRaWAN. While the Q-learning agent primarily handles the scheduling, the trust mechanism ensures that

unreliable or malicious nodes are filtered that indirectly support more effective and reliable network operations.

V. Results and Discussion

The proposed adaptive duty cycle management framework using Q-learning for LoRaWAN is analytically modeled and scripted in Python 3.10. The simulations are carried out considering a Python-based environment using Jupyter IDE/Notebook. The simulation framework considers a custom discrete-event simulation modeling in Python for LoRaWAN. It also implements a custom implementation of Q-learning, considering visualization libraries, data handling libraries, and machine learning (ML) libraries. The security module also incorporates custom filtering and blacklist logic. The simulation system configuration considers an Intel Core i5 processor with 12 GB RAM and a Windows operating system. The custom Q-learning is CPU-light for conceptual modeling, so it doesn't require GPU processing.

The experimental evaluation considered a custom discrete-event simulation model developed in Python to emulate LoRaWAN communication under varying traffic and channel conditions. The proposed framework also includes realistic parameters such as bandwidth (125 kHz), transmission power levels consistent with LoRa Class-A devices. For comparison, the popular baseline schemes such as Static (fixed duty-cycle allocation), Random (stochastic transmission slot selection), and Round Robin (cyclic slot assignment) were implemented. All metrics, such as throughput (packets/sec) and energy (mJ), are normalized averages derived from the simulation experiments. The network topology configuration is shown in Table I.

TABLE I. Simulation Configuration

Parameters	Values
Number of IoT Devices	10-100
Number of LoRa gateways	1-2
Simulation Time Step	50
No. of Episodes	1000
Coverage Radius (per Gateway)	1000 meters
Device Transmission Slots	10 slot cyclic duty schedule
Random Packet Loss Probability	10-15%
Attack Injection Rate	10% (Fake Sensor Values)
Initial Battery Level	100 Joule
$\boldsymbol{TX_{high}}$ Energy Consumption	5 units
$\boldsymbol{TX_{low}}$ Energy Consumption	2 units
$\boldsymbol{wait}$ Energy Impact	0 (only incurs delay penalty)
Learning Rate α	0.1
Discount Factor $\boldsymbol{\gamma}$	0.9
Exploration Rate $\boldsymbol{\varepsilon}$	0.1

The proposed work observes the trends from the plots after implementing the strategy of adaptive duty cycle operation in LoRaWAN. The observed trends from Q-learning is illustrated in Fig. 3. Fig. 3(a) shows that trend in cumulative reward per episode is initially low, but improves and stabilizes over episodes. This indicates the Q-learning agent in the proposed system learns effective actions to maximize successful transmission and conserve energy while also avoids unnecessary transmissions as the encouragement with higher reward in proposed Q-learning leads to better decision policy.

The analysis of the buffer size also implies that the buffers remain within mid to low ranges, which clearly shows high variability as buffer size fluctuates significantly from episode to episode. The buffer size ranges from 0-4 packets [see Fig. 3(b)]. The high-frequency fluctuations across the episode range suggest that the buffer size is dynamic and frequently changing. The buffer size is rarely remains zero or maxim indicating the system is actively transmitting and receiving data. Values between 1 and 3 dominate suggest a moderately filled buffer across the learned period. The observed variability also indicates that the system adapts dynamically to changes in network load channel availability and energy constraints. Since learning policy involves exploration using ε-greedy policy hence fluctuations are expected as the agent explores different transmission strategies. Overtime although no clear smoothing is visible, the agent avoids extreme buffer overflows or starvation indicating policy convergence. The observed variation suggests a balance between energy and throughput.

Here, Fig. 4 indicates the number of successfully transmitted packets received by the gateway at each time step. Here, the throughput values fluctuate based on agent's learned behavior and environmental dynamics such as channel congestion or blacklist filtering. The Q-learning agent gradually improves its decision-making strategy between TX_{high} , TX_{low} and $wait$ to prioritize successful transmissions when the channel is idle and the battery is sufficient. This adaptive learning process prevents excessive collision or batter drain and ensure stable throughput over time. It also shows good responsiveness and adaptive behavior with increasing number of IoT End Devices (EDs).

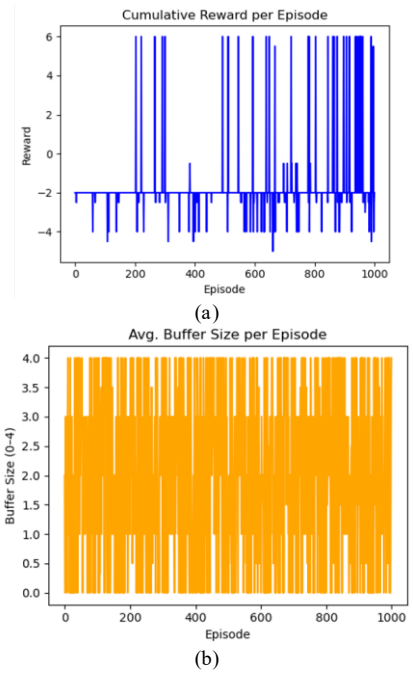


Fig. 3. Observed trends from Q-learning: a) analysis of cumulative rewards and b) average buffer size per episode.

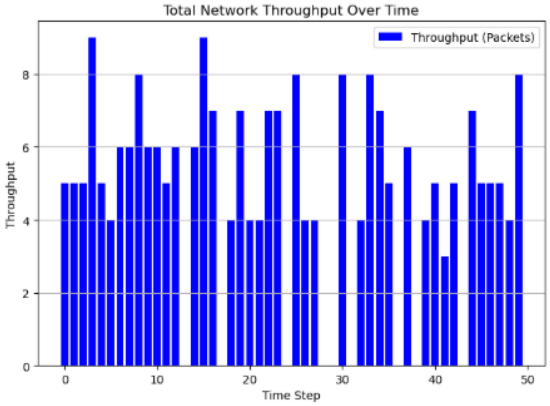


Fig. 4. Analysis of throughput in LoRaWAN.

The observed trends from Q-learning is illustrated in Fig. 5. The interpretation of the [Fig. 5(a)] plot shows ratio of packets successfully transmitted relative to total number of active devices at each time step. The proposed Q-learning based algorithm in LoRaWAN shows moderately efficient bandwidth utilization trend that clearly depicts the fact that the Q-agent balances the aggressive and conservative transmission attempts. Low utilization at certain intervals correlates with high interference (channel status busy) or conservative WAIT actions, which the agent learns to apply when risk is high or energy is low. This approach significantly helped reducing the congestion and packet loss. It offers ~65% utilization and the algorithm with this approach avoids overloading the network while maintains consistent data delivery and indicate balanced usage of resources. Fig. 5(b) captures the average battery level of all IoT devices over simulation steps. It highlights decreasing but controlled energy curve that signifies an energy-aware strategy in LoRaWAN. It also depicts that learning agent adaptively decide and favor low power transmission or defer transmission when essential as guided by the Q-learning optimal policy. The gradual descent also reflects sustainable energy usage rather than rapid depletion which verify the framework's suitability for long term operation in large-scale and constrained LoRaWAN.

The comparison of throughput for proposed Q-learning in (Fig. 6) shows that it achieves highest throughput in initial steps in LoRaWAN. Slowly it starts adapting to the environment dynamics and select transmission times that maximizes rewards. This leads to higher and smarter throughput over time. However, in random policy inconsistency is observed in the throughput outcome as it occasionally performs well by chance but lacks reliability and wastes resources due to frequent collisions or poor timing. However Round Robin offers stable but low throughput as it underutilizes idle slots and cannot adapt to traffic or channel conditions. Static duty cycle paradigms are fragile in dynamic environments and result poor reliability and wasted capacity. The Q-learning approach also occasionally yields zero throughput at specific time steps these reflect intelligent decision making such as avoiding unfavorable transmission slots or conserving energy. Here, Q-learning adapts its policy to maximize long-term throughput leading to overall superior performance despite transient dips in LoRaWAN.

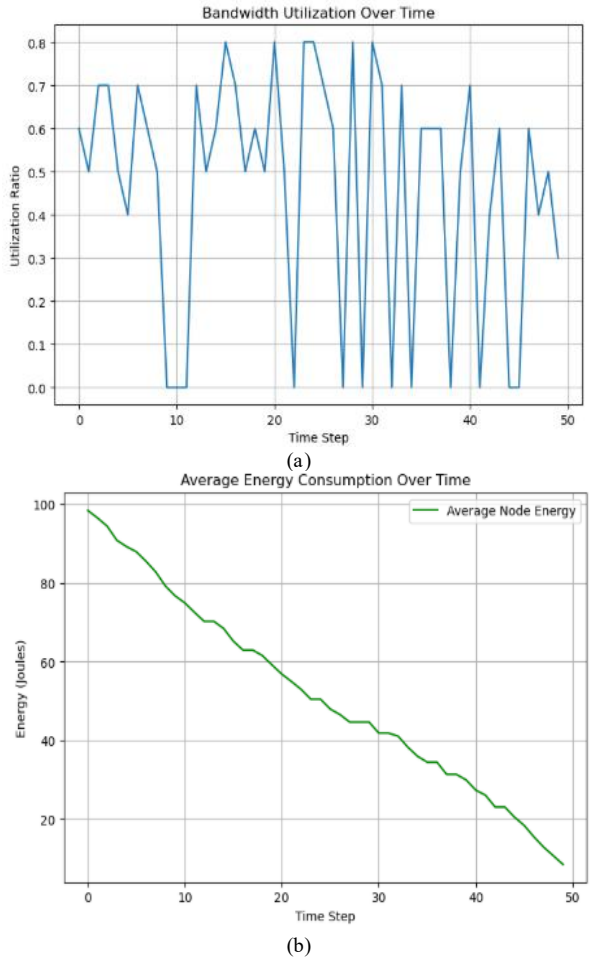


Fig. 5. Observed trends from Q-learning: a) analysis of bandwidth utilization ratio, and b) average energy consumption over time.

The delay outcome is found highly stable in the case of Q-learning strategy, as it learns from the reward feedback to choose actions (transmission slots) that minimizes delay. After brief learning phase, it converges to an efficient schedule with minimal delay. However, other methods such as Random, Round Robin and Static either transmit blindly or without adaption to repeated conflicts therefore, suffers from instability, higher delay variance (see Fig. 7).

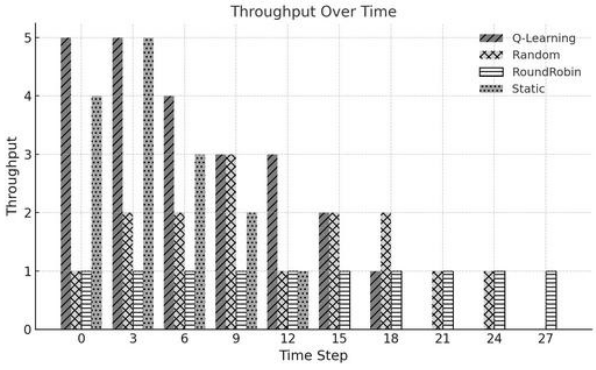


Fig. 6. Comparison of throughput with different approaches.

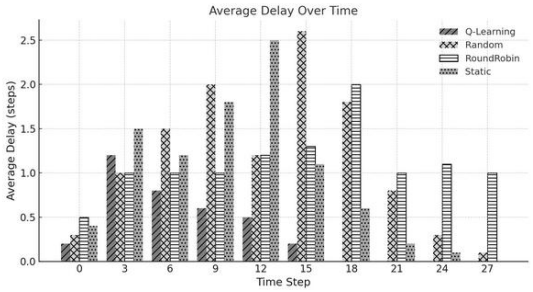


Fig. 7. Comparison of delay outcome with different approaches.

It can be seen from Fig. 8 that Q-learning rapidly explores its action space in early stages which consumes energy. However, once it learns the optimal transmission pattern then energy consumption of IoT devices significantly drops and remains stable. However, it also outperforms others by minimizing the energy use once the optimal policy is learned which static and non-adaptive schemes fail to achieve in LoRaWAN.

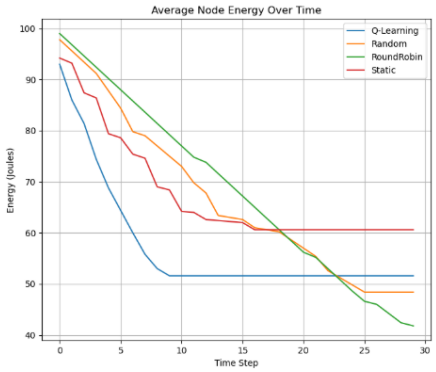


Fig. 8. Comparison of energy outcome with different approaches.

Fig. 9 reveals how many devices are being blacklisted throughout the system's operation in LoRaWAN. The blacklist count increases gradually which suggests that the security system is capable of detecting and isolating misbehaving or low-trust devices which mostly forward suspicious packets in LoRaWAN environment. Here, Q-learning plays a crucial role towards evaluating device behavior and taking penalizing actions over time. The trust-driven approach in the proposed scheme not only improves security and reliability but also contributes towards optimal throughput, lower delay and stable energy consumption.

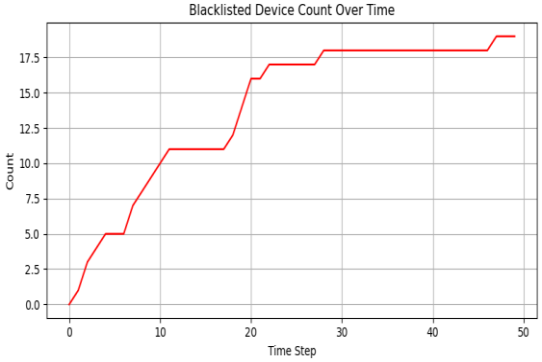


Fig. 9. Analysis of security in LoRaWAN.

VI. Conclusion

The proposed work introduces a Q-learning-based adaptive scheduling and duty cycle strategy to enhance the performance of LoRaWAN. Here, the proposed adaptive duty cycling approach incorporates Q-learning to dynamically choose the best time slot based on traffic, battery and buffer status reducing congestion and improving energy efficiency. Additionally, the unified framework also offers dynamic energy management and optimized transmission under traffic variability. An efficient channel scheduling algorithm is also proposed to enhance the bandwidth utilization, throughput and delay performance in LoRaWAN. The performance of the proposed Q-learning is also evaluated against Static, Random, and Round Robin approaches. The experimental results show that the Q-learning significantly improves the throughput performance (peak: 5 units), and also ensure lower average delay (converging to 0 within 10 steps), and improved energy efficiency (stabilizing at ~52J, ~15% better than Round Robin). It also offers a security analysis strategy that effectively identifies up to 19 malicious or underperforming devices via dynamic blacklisting. The outcome also conceptually justifies that the learning-based approach adapts well to traffic variations and security threats, outperforming static and randomized methods. The proposed framework demonstrates how reinforcement learning and lightweight edge security can jointly enhance LoRaWAN performance and ensure adaptive, energy-efficient, and secure communication in large-scale IoT deployments. In future work, we aim to extend this framework using deep reinforcement learning, multi-agent collaboration, and trust-aware reward models, with validation on more extensive LoRaWAN deployments.

References

[1] S. Devalal and A. Karthikeyan, "LoRa technology—An overview," in Proc. 2nd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA), Coimbatore, India, Mar. 2018, pp. 284–290. doi: 10.1109/ICECA.2018.8474803.

[2] N. Telagam, N. Kandasamy, and D. Ajitha, "Smart healthcare monitoring system using LoRaWAN IoT and machine learning methods," in Practical Artificial Intelligence for Internet of Medical Things, CRC Press, 2023, pp. 85–104. doi: 10.1201/9781003362725-6.

[3] 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. 1–7, Mar. 2019. doi: 10.1016/j.icte.2019.01.005.

[4] C. Gomez, J. C. Veras, R. Vidal, L. Casals, and J. Paradells, "A Sigfox energy consumption model," Sensors, vol. 19, no. 3, p. 681, Feb. 2019. doi: 10.3390/s19030681.

[5] A. Farhad, D. H. Kim, B. H. Kim, A. F. Y. Mohammed, and J. Y. Pyun, "Mobility-aware resource assignment to IoT applications in long-range wide area networks," IEEE Access, vol. 8, pp. 186111–186124, 2020. doi: 10.1109/ACCESS.2020.3029947.

[6] D. Kjendal, "LoRa-Alliance regional parameters overview," J. ICT Standardization, vol. 9, no. 1, pp. 35–46, 2021. doi: 10.13052/jicts2245-800X.9113.

[7] S. R. Borkar, "Long-term evolution for machines (LTE-M)," in LPWAN Technologies for IoT and M2M Applications, Academic Press, 2020, pp. 145–166. doi: 10.1016/B978-0-12-818880-4.00010-1.

[8] J. Haxhibeqiri, E. De Poorter, I. Moerman, and J. Hoebeke, "A survey of LoRaWAN for IoT: From technology to application," Sensors, vol. 18, no. 11, p. 3995, Nov. 2018. doi: 10.3390/s18113995.

[9] K. L. Tsai, F. Y. Leu, I. You, S. W. Chang, S. J. Hu, and H. Park, "Low-power AES data encryption architecture for a LoRaWAN," IEEE Access, vol. 7, pp. 146348–146357, 2019. doi: 10.1109/ACCESS.2019.2945801.

[10] O. B. Seller and N. Sornin, Low Power Long Range Transmitter, U.S. Patent 9,252,834, Feb. 2, 2016.

[11] C. A. Gomez, A. Shami, and X. Wang, "Machine learning aided scheme for load balancing in dense IoT networks," Sensors, vol. 18, no. 11, p. 3779, 2018. doi: 10.3390/s18113779.

[12] T. Yatagan and S. Oktug, "Smart spreading factor assignment for LoRaWANs," in Proc. IEEE Symp. Comput. Commun. (ISCC), Barcelona, Spain, Jun. 2019, pp. 1–7. doi: 10.1109/ISCC47284.2019.8969645.

[13] S. U. Minhaj et al., "Intelligent resource allocation in LoRaWAN using machine learning techniques," IEEE Access, vol. 11, pp. 10092–10106, 2023. doi: 10.1109/ACCESS.2023.3241097.

[14] S. Cui and I. Joe, "Collision prediction for a low power wide area network using deep learning methods," J. Commun. Netw., vol. 22, no. 3, pp. 205–214, Jun. 2020. doi: 10.1109/JCN.2020.000016.

[15] L. Liu, Y. Yao, Z. Cao, and M. Zhang, "DeepLoRa: Learning accurate path loss model for long distance links in LPWAN," in Proc. IEEE INFOCOM, Vancouver, BC, Canada, May 2021, pp. 251–260. doi: 10.1109/INFOCOM42981.2021.9488831.

[16] S. Lee, J. Lee, J. Hwang, and J. K. Choi, "A novel deep learning-based IoT device transmission interval management scheme for enhanced scalability in LoRa networks," IEEE Wireless Commun. Lett., vol. 10, no. 11, pp. 2538–2542, Nov. 2021. doi: 10.1109/LWC.2021.3104133.

[17] S. I. A. Elkarim et al., "Deep learning based joint collision detection and spreading factor allocation in LoRaWAN," in Proc. IEEE 42nd Int. Conf. Distributed Comput. Syst. Workshops (ICDCSW), Bologna, Italy, Jul. 2022, pp. 187–192. doi: 10.1109/ICDCSW55796.2022.00054.

[18] R. M. Sandoval, A. J. Garcia-Sanchez, and J. Garcia-Haro, "Optimizing and updating LoRa communication parameters: A machine learning approach," IEEE Trans. Netw. Serv. Manag., vol. 16, no. 3, pp. 884–895, Sep. 2019. doi: 10.1109/TNSM.2019.2934601.

[19] N. Aihara, K. Adachi, O. Takyu, M. Ohta, and T. Fujii, "Q-learning aided resource allocation and environment recognition in LoRaWAN with CSMA/CA," IEEE Access, vol. 7, pp. 152126–152137, 2019. doi: 10.1109/ACCESS.2019.2947754.

[20] M. Jouhari, N. Saeed, M. S. Alouini, and E. M. Amhoud, "A survey on scalable LoRaWAN for massive IoT: Recent advances, potentials, and challenges," IEEE Commun. Surveys Tuts., vol. 25, no. 3, pp. 1841–1876, 2023. doi: 10.1109/COMST.2023.3244018.

[21] F. Adelantado, X. Vilajosana, P. Tuset-Peiro, B. Martinez, J. Melia-Segui, and T. Watteyne, "Understanding the limits of LoRaWAN," IEEE Commun. Mag., vol. 55, no. 9, pp. 34–40, Sep. 2017. doi: 10.1109/MCOM.2017.1600613.

[22] T. Deng, J. Zhu, and Z. Nie, "An improved LoRaWAN protocol based on adaptive duty cycle," in Proc. IEEE 3rd Inf. Technol. Mechatronics Eng. Conf. (ITOEC), Chongqing, China, Oct. 2017, pp. 1122–1125. doi: 10.1109/ITOEC.2017.8320815.

[23] A. Maleki, H. H. Nguyen, E. Bedeer, and R. Barton, "A tutorial on chirp spread spectrum modulation for LoRaWAN: Basics and key advances," IEEE Open J. Commun. Soc., 2024. doi: 10.1109/OJCOMS.2024.3384025.

[24] D. Kalugina and A. Pastukh, "Evaluation of the feasibility of implementing satellite IoT in the terrestrial IoT frequency bands of 868 MHz and 915 MHz," in Proc. Int. Sci. Tech. Conf. Modern Comput. Netw. Technol. (MoNeTeC), Oct. 2024, pp. 1–6. doi: 10.1109/MoNeTeC60266.2024.10406290.

[25] X. Yang, E. Karampatzakis, C. Doerr, and F. Kuipers, "Security vulnerabilities in LoRaWAN," in Proc. IEEE/ACM 3rd Int. Conf. Internet-of-Things Design and Implementation (IoTDI), Apr. 2018, pp. 129–140. doi: 10.1109/IoTDI.2018.0001.

[26] P. Yu, F. Zhou, X. Zhang, X. Qiu, M. Kadoch, and M. Cheriet, "Deep learning-based resource allocation for 5G broadband TV service," IEEE Trans. Broadcast., vol. 66, no. 4, pp. 800–813, Dec. 2020. doi: 10.1109/TBC.2020.3004470.

[27] S. U. Minhaj et al., "Intelligent resource allocation in LoRaWAN using machine learning techniques," IEEE Access, vol. 11, pp. 10092–10106, 2023. doi: 10.1109/ACCESS.2023.3241097.

[28] S. Lavdas, N. Bakas, K. Vavousis, W. El Hajj, and Z. Zinonos, "Evaluating LoRaWAN network performance in smart city environments using machine learning," IEEE Internet Things J., 2025. [Early Access]. doi: 10.1109/JIOT.2025.339412.

[29] M. Alkhayyal and A. M. Mostafa, "Enhancing LoRaWAN sensor networks: A deep learning approach for performance optimizing and energy efficiency," Computers, Materials & Continua, vol. 83, no. 1, 2025. doi: 10.32604/cmc.2025.046261.

[30] A. Farhad and J. Y. Pyun, "LoRaWAN meets ML: A survey on enhancing performance with machine learning," Sensors, vol. 23, no. 15, p. 6851, 2023. doi: 10.3390/s23156851.

Real-Time Multi-Scale Object Detection in Surveillance Using Hybrid Transformer Architecture

Roshan D Suvaris[1], Rahul Suryodai[2], S. Narayanasamy[3], Dr. Aanandha Saravanan[4], Raman Kumar[5], Dr. P N V Syamala Rao M[6], Elangovan Muniyandy[7]

Asst. Professor, Nitte (Deemed to be University), NMAM Institute of Technology (NMAMIT), Nitte, India [1]
Senior Data Engineer (Data Governance, Data Analytics: Enterprise Performance Management, AI&ML), USA[2]
Department of Computer Science and Engineering, J. J. College of Engineering and Technology, Tiruchirappalli, Tamilnadu, India[3]
Professor, Department of ECE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India[4]
University School of Mechanical Engineering, Rayat Bahra University, Mohali, India[5a]
Faculty of Engineering, Sohar University, Sohar, Oman[5b]
Assistant Professor, Department of CSE, SRM University AP, Amaravati, Andhra Pradesh, India – 522240[6]
Department of Biosciences-Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai - 602 105, India[7a]
Applied Science Research Center, Applied Science Private University, Amman, Jordan[7b]

***Abstract*—Real-time surveillance systems require accurate and efficient object detection to ensure safety and situational awareness. Existing methods, such as YOLOv5 and Vision Transformer-based detectors, often struggle to reliably identify small, distant, or occluded objects while maintaining real-time inference, limiting their applicability in complex surveillance environments. To address these challenges, this study proposes PRISM, a hybrid Transformer–YOLOv8 framework that integrates fast local feature extraction with global contextual refinement. The method introduces two novel components: i) a Context-Aware Feed Forward Network (CA-FFN) within the Vision Transformer (ViT), which dynamically weights channel features to reduce redundancy and enhance global context modeling, and ii) Cross-Scale Attention Skip Connections (CSASC) for selective fusion of multi-scale YOLOv8 and ViT features, improving detection of small or occluded objects. The model is implemented in PyTorch and trained on a comprehensive surveillance dataset consisting of pedestrians, vehicles, bicycles, bags, and miscellaneous objects. Experimental evaluation demonstrates that PRISM achieves 96% accuracy, a significant improvement of ~4–5% over baseline methods, with robust performance across all object categories. Key performance indicators verify the reliability of the model to real-time usage, and the lightweight design makes it edge deployable. These findings imply that PRISM can be used to provide a speed-accuracy balance in a complex and dynamic setting, which is more efficient than the current methods. The study also notes the partial extensions, such as the incorporation of multi-sensors and continuous video streams to do time modeling as an extension, which will offer a good base to the next-generation intelligent surveillance systems.**

***Keywords*—Real-time object detection; hybrid transformer–YOLOv8; Context-Aware Feed Forward Network (CA-FFN); Cross-Scale Attention Skip Connections (CSASC); surveillance video analytics; multi-scale feature fusion**

I. Introduction

Real-time surveillance has turned into an inseparable component of the contemporary governance of the city, social security and management of vital infrastructure [1]. Since the idea of smart cities has entered a new phase and the necessity of automated monitoring devices [2] is growing, the ability to properly distinguish and classify objects in dynamic and complicated environments has received primary importance [3]. These traditional computer vision algorithms have been largely reliant on handcrafted features that are sensitive to varying light conditions, occlusion, and size, and hence limit their application in real-world surveillance [4]. With the emergence of deep learning (DL) algorithms, and in particular convolutional neural networks (CNNs) [5], object detection has gotten a significant boost due to automated feature extraction and strong recognition features. The You Only Look Once (YOLO) family of CNN-based [6] detection methods has become one of the most popular frameworks in real-time detection because it operates in a single stage, and all images are processed in a single pass, having an impressive inference speed [7]. However, the YOLO-based models are mostly centered on the local spatial information, disregarding the contextual relationships at a large scale, which may worsen the performance of the model on the detection of small, hidden, or distant objects [8]. ViTs, on the other hand, make use of self-attention mechanisms to learn global dependencies [9], which give a better representation of more complex scenes and greater detection of difficult-to-detect object classes. However, ViTs are computationally expensive and latency-prone, which makes them less ideal for real-time execution on edge devices [10]. Hybrid models using CNN backbones and transformers have also been studied, but most existing methods are inclined to excessive computational cost or non-optimal multi-scale feature integration, which leads to a trade-off between efficiency and accuracy [11], [12].

To address these challenges, this study presents PRISM, a new hybrid system of scalable real-time monitoring of surveillance objects. PRISM combines the property of extracting features of objects within a short time frame used in YOLOv8 with a lightweight ViT module, with an addition of a Context-Aware Feed-Forward Network (CA-FFN) that boosts the network's understanding of the objects in their entirety. This is done using a Cross-Scale Attention Skip Connection (CSASC) mechanism that is effective in producing local and global features at multiple scales without compromising speed and accuracy. The framework integrates powerful preprocessing, data augmentation, and CIoU-based optimization to achieve maximum performance in the detection of various categories of objects such as pedestrians, vehicles, bicycles, and miscellaneous objects. PRISM aims to work around the shortcomings of the current method, combining the strengths of YOLOv8 and ViTs to provide a better detection of small, occluded, and distant objects, preserving the performance of real-time detection. The proposed framework would be applicable to both edge and cloud deployments and would be flexible, scaled, and robust in the present-day surveillance applications.

A. Research Motivation

The increased complexity of urban observation contexts (i.e., occlusion, changing light, and object size variability) requires detection systems to achieve high accuracy while meeting real-time demands [13]. Existing models rely on speed or are not generalizable when faced with challenging situations, which creates the need to develop a generalizable, scalable hybrid approach to robust object recognition.

B. Significance of the Study

This study addresses important gaps in present-day surveillance object detection by utilizing multi-scale contextual learning and fast inference. The proposed framework improves the detection of small, occluded, and distant objects to collect reliable data for a better and more effective public safety, traffic management, and infrastructure protection. Lastly, this solution is deployable across edge and cloud devices.

C. Recent Innovation and Challenges

Recent advancements, such as transformer-based models and hybrid CNN-transformers, have increased the quality of representation of global features. However, there are still difficulties in achieving a balance between computational efficiency, multi-scale feature fusion, and real-time operation. Integrating lightweight attention mechanisms effectively is still a primary challenge to implement practical, scalable surveillance solutions in changing, dynamic environments.

D. Key Contributions

- Hybrid Framework Design: Development of PRISM, a scalable object detection framework that combines fast local feature extraction with global contextual refinement for robust real-time surveillance.
- Novel Fusion Mechanism: Innovation of CSASC to improve multi-scale feature fusion while ensuring computational efficiency.
- Transformer Optimization: Addition of a CA-FFN to the lightweight ViT module to enhance global feature learning on small, occluded, and faraway objects.
- Robust Preprocessing Pipeline: Use of a systematic data cleaning, augmentation, normalization, and adaptive training approach to achieve model generalization in various surveillance environments.

The remaining section of this study will follow the below organization. In Section II, prior research in real-time detection in the transfer application will be explored, examining findings and restrictions of YOLO and Transformer-based models. In Section III, the most notable challenges in surveillance object detection are identified. These are small object recognition, occlusions, and computational efficiency. The proposed Scalable Transformer-YOLO Model methodology is discussed in Section IV and includes preprocessing of the dataset, architecture of the model, and training methodologies. Section V presents results from experimentation, performance metrics, and comparisons to other conventional models. Section VI concludes with a few conclusions, along with some major discussion and potential future improvement work for the real-time surveillance detection system outcomes.

II. Related Works

Ouyang [14] was proposed as a hybrid system, which combines the concepts of DETR and YOLO to become more robust and more accurate. It is a two-stage pipeline architecture in which the bounding boxes of objects were predicted by YOLO extremely quickly. Then, to add high precision transformer-based model refinement was employed. On the COCO dataset, DEYO has achieved an achievement of 52.1 AP and has surpassed the traditional YOLO models in both detection and performance. The combination of YOLO and transformers assists in the local and global contexts, which progresses the object recognition in difficult scenes. Although DEYO improved the accuracy, it had a negative response as it took a long time to converge and was computationally too costly to use in real-time. The principal disadvantage was its accompanying downside of preparing a large amount of training data and adjusting hyperparameters, consuming resources. Additionally, the transformer-based refinement added even more latencies, restricting its capacity for real-time surveillance applications.

Song et al. [15] proposed ViDT, a novel object detection model based solely on the ViTs, which has subsequently improved detection accuracy. ViDT, in contrast to earlier CNN-based models, fabricates a mechanism for retrieving long-range image dependencies, which has improved object recognition. The model has displayed an AP of 49.2 on the COCO dataset; thereby, it outperforms conventional object detection models. The capacity to learn contextual relationships was one of the main benefits of this model and proved useful in situations when objects were interacting. It is, however, a resource-consumptive and, therefore, not applicable to near-real-time edge solutions that require improved resource utilization. The other weakness is that it is based on training on large datasets, and this usually translates to the transformers requiring much more labeled data in order to perform optimally. The necessity to have lightweight models of transformers that would create a

reasonable balance between accuracy and efficiency in computation has been mentioned in the work.

Wang et al. [16] suggested Mamba YOLO, a state-of-the-art model for object detection, which incorporated the use of SSMs in the YOLO to enhance efficiency and accuracy. Mamba YOLO, which makes use of SSMs in the continuity of its one-step thinking process, allows tracking of objects through time. In this way, it greatly aids by reducing the complexity of computation and hence completely enables the applications of Real-time in autonomous systems and surveillance. The model demonstrated huge progress on the COCO and VOC datasets with large improvements compared to classical YOLO versions on both precision and recall. Another weakness was the inability to perform better in extreme light conditions because objects with poor contrast were hard to notice. Nonetheless, Mamba YOLO had potential for making a valuable contribution to the real-time detection area, given the blend of YOLO's speed with SSMs' sequential page processing abilities.

G. and B. [17] created YoloTransformer-TransDetect, a hybrid in that their framework merges YOLO with transformer-based attention mechanisms to implement defect detection. The sequence included first using YOLO to quickly identify defects, followed by the transformer module that worked to extract features at a finer level of grain to improve accuracy. There was better performance of the model in detecting defects in steel tubes, as well as an increase in the detection rates. The technique was not able to work with extremely complicated or visually similar defect patterns in a single pass and had to undergo additional post-processing. A tiny defect and occlusion, one of the common problems in industry, was classically privileged when the method could reliably, if not efficiently, identify. The research found that some feature learning and localization were realized using transformer-based techniques along with YOLO. Its operation efficiency, however, was questioned, and restricted its use in real-time. The future research undertakes the aim of optimization of transformer block to offer convenience in shortcomings.

Li, Yan, and Shi [18] introduced PP-YOLOE, a new version of YOLO for object detection, which comes with a multi-scale attention mechanism. The adaptive scaling method improves feature extraction to better specify fine distinction details in every scale of any object. Application of these mechanisms significantly enhanced the total accuracy of detection on the COCO data set, mainly on small objects, thus are highly overlooked in previous traditional YOLO models. This network, with the help of attention-based techniques, was rather a skilled object localizer in cluttered and occluded scenes, where it provided fewer false positives. Latency can limit applications in which real-time decision-making is required as video surveillance or even autonomous driving. It is also suggested in the study that lightweight models of transformer should be followed up in order to achieve a good speed-accuracy trade-off.

Su et al. [19] came up with a different model of breast mass detection as they divided the mass into LOGO and applied the strengths of YOLOv5 to it. The algorithm is a two-step process and involves both segmentation and detection of breast masses. First, the image is segmented and cropped with the help of a YOLOv5L6 model on high-resolution mammograms; then, the image and the area with breast masses are processed by various global and local transformer branches; finally, they are combined to yield the final segmentation output. Although these are encouraging figures, integration of transformers has computational difficulties, and it is not an easy task to implement it in a regular hospital system that has low computing capabilities. The study revealed that fusing transformers with YOLO improves boundary refinement and lesion detection but requires excessive data augmentation for generalization onto other imaging datasets.

Shah and Tembhurne [20] developed Defect Transformer, a hybrid architecture utilizing transformers, set to leverage surface defects in industrial applications, which is based on attention-based feature extraction for localization and classification of defects, as well as other aspects. The model was reported to be more accurate in detecting microscopic defects compared to the traditional CNN models. It is however, limited in application when it comes to real-time monitoring due to extremely high levels of computational complexity. This study demonstrated that perceiving the world context can be useful in bettering the process of defect classification. Although it can have multiple applications, there are major limitations to its implementation in real-time applications because of the memory and processing requirements of embedded systems.

Shang et al. [21] presented the Defect-Aware Transformer Network, a deep learning framework designed for intelligent defect detection in modern industrial applications. The model integrated a self-attention mechanism to improve feature representation, which helps to detect those defects that were considered subtle and fine-grained, as is often missed with traditional convolutional models. This was in addition to the mix that was the transformer-based feature extraction and the conventional DL methodologies, which offered good learning of both local and global defect patterns. The experimental tests have demonstrated that the model resulted in a significant improvement of the trade-off between the accuracy and robustness of the classification, in particular, against defects with low contrast and textured backgrounds. In spite of these strengths, the authors also mentioned that their method is computationally expensive and thus not practical in real-time use in the automation of industry.

Current object detection studies highlight hybrid models, where YOLO is combined with transformers to make a balance between speed and precision. Hybrid methods enhance accuracy, but are computationally expensive. Pure transformer-based models detect long-range dependencies and contextual associations, but depend on large datasets and intensive resources. Some models combine sequential modules or multi-scale attention to achieve real-time efficiency and identify fine details. Although refinements based on transformers improve overall detection performance and localization, they also increase latency and computational requirements, which reduces their feasibility in real-world applications such as surveillance, autonomous systems, and industrial defect detection.

III. Problem Statement

Real-time object detection forms are one of the core components of surveillance systems where accuracy, flexibility, and efficiency are of equal significance. YOLO-based models are appreciated due to their lightweight architecture and fast inference that allows them to be deployed in edge cases and real-time systems. However, they do not properly identify small, distant, or hidden objects that usually appear in crowded and dynamic surveillance conditions [16]. Transformer-based models, on the other hand, are good at capturing global spatial context and enhance the quality of detection, but are costly to compute, and thus are restricted in their use in resource limited or real-time environments [22]. Hybrid architectures that integrate YOLO with transformer-based architectures have been reached to overcome those drawbacks, but they are generally accompanied by additional issues such as slow convergence, large memory complexity, and scalability constraints in the field environment. These drawbacks leave a research gap that hinders the development of smart surveillance. This is dealt with in the present work by the development of an optimized hybrid framework that trades accuracy, contextual learning, and efficiency of inference both in edge and cloud-based environments.

IV. Proposed PRISM Framework for Real-Time Object Detection

The suggested PRISM framework is modelled as the powerful and scalable object-detecting built-in surveillance system in real-time that is able to take into account the challenges of crowded scenes, changing illumination, partiality, and small sizes of items. The architecture is a hybrid between the convolutional and transformer-based architecture to trade local detail speech and global contextual reasoning. YOLOv8 is essentially the backbone network, and is effective at extracting multi-scale local features giving spatial precision and fast inference, which are essential to real time implementation. Following the extraction step, a lightweight ViT fine-tunes them by capturing long-range dependencies and high-level semantic relationships across the scene. The ViT component is enhanced with a CA-FFN to perform efficient channel-wise recalibration of the features as well as minimizing redundancy in the output. In order to fill the gap between local spatial information and global context, a new CSASC mechanism is added for enabling the proper merging of YOLOv8 features with ViT-refined results. The combined detection head subsequently executes bounding box regression and classification, which is optimized using CIoU-based loss and adaptive learning policies. Overall, this hybrid architecture provides a high-accuracy real-time system that is still computationally lightweight and well-suited for cloud and edge deployments in surveillance systems.

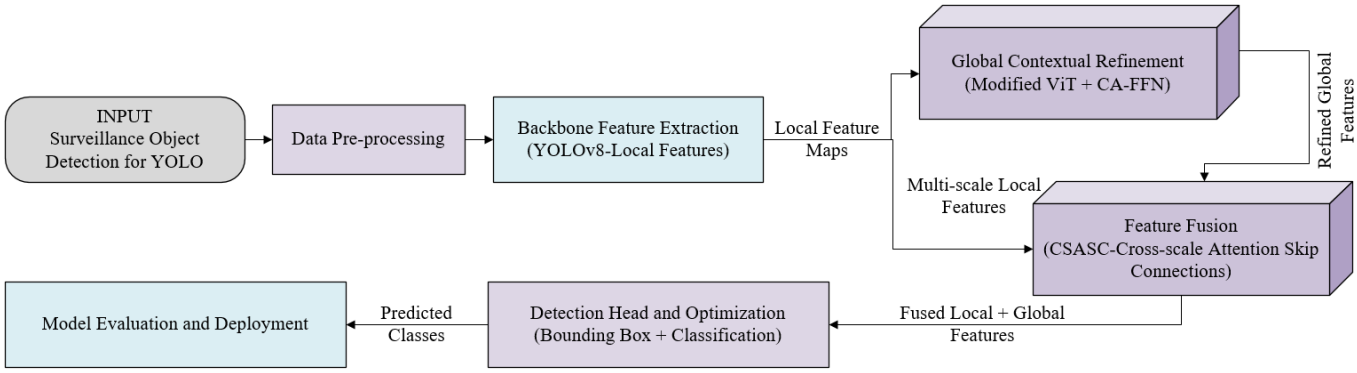


Fig. 1. Proposed PRISM framework.

Fig. 1 shows the process used to detect by the proposed system. The preprocessing of input video streams is followed by feature extraction by means of the YOLOv8 backbone and a lightweight ViT. The extracted features are then used to perform object detection and classification after which the outputs undergo post-processing to give a higher caliber. Evidence is formed to evaluate the quality of detection. The framework is then capable of implementing real-time decision-making or deployment on edge and cloud platforms, which guarantees scalability, efficiency, and flexibility in various settings of surveillance

A. Dataset Overview

The "Surveillance Object Detection Dataset for YOLO" [23] from Kaggle aims to create a training and test dataset with labeled training images specifically for surveillance real-time object detection applications. Approximately, 1,000 annotated images were taken based on real-world surveillance footage, and they depict vehicles and pedestrians, as well as objects of interest, the images are annotated with bounding boxes in the YOLO format that can be used with YOLOv8 models. In every annotation file, there are the class IDs, and coordinates of the bounding box in normalized form (0 to 1) which implies that it is resolution-independent. The dataset is most suitable in the real-time security applications including smart monitoring, anomaly detection, and traffic monitoring whereby strong object detection is a must. It has a great foundation with its formal annotation format and core object types that offer a strong and realistic platform to train YOLOv8 based surveillance models with better accuracy and reliability. The dataset parameters were given in Table I.

TABLE I. DATASET ATTRIBUTES

Attribute	Description	Attribute
Total Images	~1,000 images (varies based on dataset version)	Total Images
Image Resolution	Varies; normalized to 640×640 pixels	Image Resolution
Object Classes	Vehicles, pedestrians, suspicious objects	Object Classes
Annotation Format	YOLO format (TXT files with bounding box details)	Annotation Format
Data Source	Surveillance camera feeds	Data Source

B. Dataset Preprocessing

The input data required for the YOLOv8 model must be acquired using preprocessing techniques for higher precision, generalization, and stability in real-world applications of surveillance. The preprocessing techniques include data cleaning, conversion of format, data augmentation, normalization, and data splitting. These are all important in improving the dataset and the performance of the model.

1) Data cleaning: Cleaning data involves finding and fixing errors, inconsistencies, gaps, or duplicate entries in a dataset. It is important for verifying the accuracy, consistency, and reliability, all of which are required in order to construct effective and trustworthy machine learning models.

a) Identification and removal of corrupt and damaged images: Images that have been corrupted by the files, have an encoding problem, or have been incompletely downloaded are filtered out of the dataset before training [24]. Corrupt files lead to loading errors, higher overhead of computation, or non-batch processing, and therefore instability during model training. The images in the dataset that cannot be read might cause the model to drop out some training examples, and this can also lead to inconsistency in the learning of features and disrupt gradient changes.

Corrupt images are those that fail the image integrity checks. It ensures that all files are complete, well-formatted, and can be accessed. Such images should be deleted, particularly for programs, such as DL models like YOLOv8, where model performance can degrade due to missing or incomplete data. Suppose I_c refers to the number of corrupt images, and I_t to the total number of images in the dataset. The percentage of corrupt image removal can be calculated as in Eq. (1):

$$P_c = \left(\frac{I_c}{I_t}\right) \times 100 \quad (1)$$

b) Filtering mislabelled or empty annotations: Every image in the dataset should ideally have a YOLO annotation file containing bounding box coordinates and class labels. If an annotation file is missing, empty, or incorrectly labeled, the affected images will be adjusted or removed accordingly. Mislabelled annotations could lead to incorrect detections, misclassifications, and an increase in the model's false positive rate, thereby lowering overall detection accuracy.

c) Identification and elimination of duplicate images: Duplicate images lead to bias in the dataset, whereby the model overfits to the frequently seen samples. It follows when the model picks redundant sonic patterns and reduces its generalization ability to unseen test data of the surveillance system. Duplicates are mostly generated if datasets were joined, for example, in cases of automatic image gathering into surveillance systems, or data augmentation errors.

As in Table II, the removal of these problematic images makes way for a better dataset, thereby leading to improved model efficacy, lesser false detection, and improved generalizations in real surveillance environments. This refined dataset is now improved and made apt for preprocessing, augmenting, and training for the YOLOv8 model.

TABLE II. DATA CLEANING OPERATIONS

Issue Identified	Issue Identified	Percentage of Dataset (%)
Corrupt/Damaged Images	120	2.40%
Mislabeled/Empty Annotations	340	6.80%
Duplicate Images	215	4.30%
Total Removed	675	13.50%

2) Format conversion: To ensure conformity, all the images are also transformed into JPEG (.jpg) or PNG (.png) file format. The bounding box annotations are also YOLO format, representing the objects by normalized coordinates in the image frame. The normalizing equation of the bounding box is in Eq. (2):

$$\mathrm{x}' = \frac{x}{W}, \mathrm{y}' = \frac{y}{H}, \mathbf{w}' = \frac{w}{W}, \mathrm{h}' = \frac{h}{H} \quad (2)$$

To make the model more robust and diverse in terms of data, augmentation is done. These changes enable the model to train to identify items in many different situations of low light, motion blur and occlusions.

3) Data augmentation: In order to increase the dataset diversity and robustness of the model, augmentation is applied. Such transformations allow the model to learn to recognize objects in various conditions ranging from low light, motion blur, and occlusions.

The augmentations are used randomly to prevent overfitting, thereby, preserving model generalization in real-world surveillance environments, as provided in Table III.

TABLE III. AUGMENTATION TECHNIQUES AND THEIR EFFECTS

Augmentation Technique	Effect on Model Performance
Scaling (Zoom in/out)	Helps recognize objects at varying distances
Rotation (±15°)	Improves robustness to different camera angles
Flipping (Horizontal)	Enhances ability to detect objects in flipped views
Contrast Adjustment	Compensates for poor lighting conditions
Gaussian Noise Addition	Improves robustness to real-world noise

4) Normalization: This process of normalizing pixel intensities is highly significant as preprocessing and allows making the training of the model stable and efficient. The 0 to 255 raw pixel values in DL typically cause certain wild variation in the weight update in back propagation. These variations can lead to the unstableness of gradient descent, sluggish convergence, and model optimization. In order to solve these issues, pixel intensity is modified, and the intensity values are scaled within a fixed range between 0 and 1 in order to ensure a uniform intensity distribution across all images for the model to learn.

This enables the model to learn features more robustly by diminishing variances in brightness and contrast across the various images. It also allows for faster convergence that will help mitigate the effects of vanishing and exploding gradients. Generally, the normalization can be mathematically expressed, as in Eq. (3):

$$\mathrm{I}' = \frac{\mathrm{I} - I_{min}}{I_{max} - I_{miin}} \tag{3}$$

where, the original pixel value is denoted by I, while I_{min} and I_{miin} represent the minimum and maximum pixel intensities, typically 0 and 255, respectively. The normalized pixel value, I', is scaled within the range [0, 1]. This normalization ensures that the input distributions across all images are consistent, enabling the model to train more effectively. As a result, the model achieves better generalization, leading to improved accuracy in real-time object detection tasks.

5) Dataset splitting: As soon as preprocessing is done, the dataset is separated into orderly and effective subsets: training, validation, and testing for testing model performance in a balanced way. A balanced separation of the dataset is very crucial to avoid overfitting, optimized the learning of the model, and accurate evaluation of unseen data. The above dataset is split with reference to the standard splitting into 70% of the entire data for training, and allows learning the shapes, object features, and positions of the bounding box,20% for validation to ensure hyperparameter tuning and performance monitoring during the training process. The validation dataset helps in ensuring that the model generalizes well while protecting against overfitting onto the training dataset.10% set aside for testing to give unbiased feedback about the model's detection accuracy on data it has not seen. The details of the distribution of this dataset into different test sets are shown in Table IV.

TABLE IV. DATASET PARTITIONING

Dataset Partition	Number of Images	Percentage (%)
Training Set	700	70%
Validation Set	200	20%
Test Set	100	10%

C. Feature Extraction

The YOLOv8 is used in the PRISM framework as the base of rapid local feature detection in the frames of surveillance video. The processing of every frame is done separately by convolutional layers of the YOLOv8 that identify low- and mid-level spatial features, including edges, textures, and contours of objects. These characteristics are essential in the detection of small, remote or partially obscured objects that are major problems in real-time video surveillance images. The architecture of YOLOv8 produces multi-scale feature maps, with multi-resolution detection heads which find objects proficiently. This feature enables precise identification of large, clearly defined objects, including vehicles, and smaller, less obvious objects, such as bicyclists or pedestrians. Formally, the extraction at each layer is given as Eq. (4):

$$F_l = \sigma(W_l * F_{l-1} + b_l) \tag{4}$$

where, F_l is feature map at layer l, W_l and b_l are convolutional kernel and bias, $*$ is convolution, and σ is the activation function. This allows YOLOv8 to obtain spatial and contextual information in every frame of the video due to the hierarchical feature map aggregation. The heads of detection make predictions of bounding boxes and class likelihoods at scales of different sizes at detection time, where larger objects are more sensitive.

YOLOv8 is a real-time, high-throughput spatial fidelity sequentially processing in the case of video datasets. The learned feature maps are subsequently fed to the small ViT, which learns the long-range dependencies and the intrinsic frame contextual relations. Such a mix is what guarantees that PRISM can guarantee the good performance of detecting the dynamic scenes and can handle the application of support to the use of occlusions, motion blur, and varying object sizes without reducing its real-time performance.

D. Global Contextual Refinement

After the local feature extraction stage, which is the YOLOv8, the PRISM framework adopts a lightweight ViT to refine the global context, i.e., long-range dependencies across the extracted robustly feature maps. Unlike the existing ViTs, the proposed architecture presents a novel addition to the traditional FFN layer with the CA-FFN that offers a higher efficiency level, less redundancy of features, and better edge correctness. The modified ViT framework is given in Fig. 2.

The CA-FFN functions by initially performing a regular linear transformation on the input feature $X \in \mathbb{R}^{N \times D}$, where N represents the number of tokens extracted from the YOLOv8 feature map, and D is the dimension for embeddings. Rather than a simple MLP, the transformed feature is subject to channel-wise attention, selectively amplifying informative features while suppressing irrelevant or redundant patterns:

$$A = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right), \quad X' = AV \tag{5}$$

In Eq. (5), Q, K, V are the query, key, and value matrices from X, and d_k is the scaling factor. The attention output $X'$is next fed into dynamic feature gating, where a learnable gating vector $G \in \mathbb{R}^D$ is used to control the flow of information, given in Eq. (6):

$$X_{CA-FFN} = G \odot GELU(X'W_1 + b_1)W_2 + b_2 \tag{6}$$

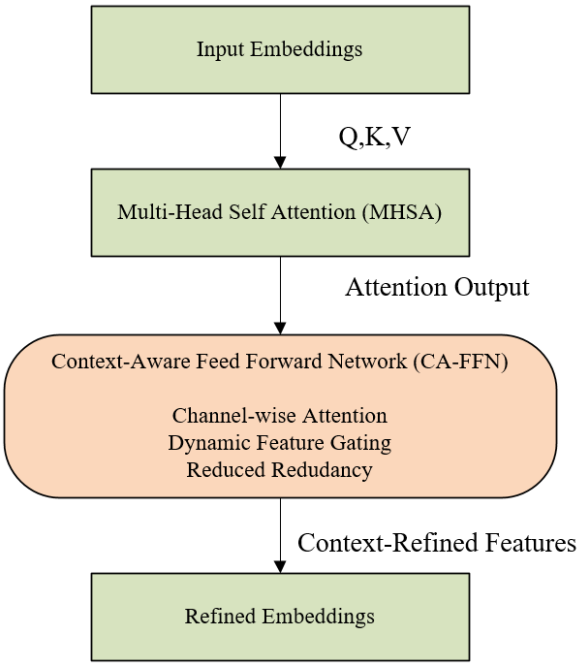


Fig. 2. Global contextual refinement.

where, W_1, W_2 and b_1, b_2 are the trainable weights and biases of the FFN, $\odot$ represents element-wise multiplication, and GELU applies non-linear activation. The gating mechanism enables the network to suppress less useful channels and concentrate computational resources on important features, leading to a significant improvement in efficiency, particularly for edge deployment applications with limited resources.

The PRISM framework, through the inclusion of CA-FFN, makes sure that global contextual information is narrowed without over-inflating computational complexity. The novelty can be easily stated: instead of the traditional ViTs, the FFN is substituted by the context-aware, attention-driven, gated module, which is, at the same time, more effective in terms of feature representation, redundancy, and lightweight architecture to fit into real-time video surveillance. The given step is essential to make sure that the fused YOLOv8 features have contextualized information, increasing the ability to find small, obscured, or visually challenging objects without making real-time performance worse.

E. Feature Fusion Mechanism

The PRISM framework uses representations by taking local feature representations provided by YOLOv8 and contextually enriched global feature representations provided by the CA-FFN-enhanced ViT, integrating them with a Feature Fusion Mechanism. Conventional skip-connections only concatenate, or add, feature maps across layers, which can tend to dilute salient information and cannot prioritize important multi-scale cues that are necessary to detect small, distant, or occluded objects. In order to address this limitation, the proposed framework includes a Cross-Scale Attention Skip Connection (CSASC), a new selective integration feature; its architecture is given in Fig. 3.

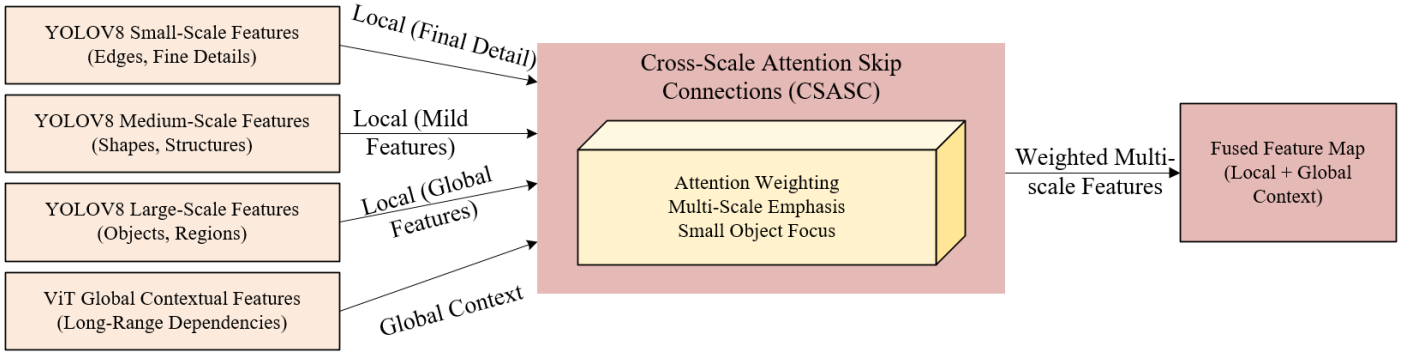


Fig. 3. CSASC fusion mechanism.

The CSASC functions by initially aligning the spatial sizes of YOLOv8 feature maps $F_{YOLO} \in \mathbb{R}^{H_1 \times W_1 \times C_1}$ and refiner-ViT features $F_{ViT} \in \mathbb{R}^{H_2 \times W_2 \times C_2}$ through bilinear interpolation and token reshaping. Then, attention maps for every scale are calculated to weigh the significance of various resolution features, as given in Eq. (7):

$$\alpha = Softmax\big(Conv_{1\times1}(F_{YOLO})\big) + Conv_{1\times1}(F_{ViT}) \quad (7)$$

$$F_{fused} = \alpha \odot F_{YOLO} + (1-\alpha) \odot F_{ViT} \quad (8)$$

In Eq. (8), α is the attention-weighting tensor derived from attention, $\odot$ is element-wise multiplication, and $Conv_{1\times1}$ is used to reduce channel dimensions for computational efficiency. This attention-based fusion allows for more attention to features with larger contexts, such as small objects, far-away vehicles, or partially occluded pedestrians, and less attention to regions that do not contain informative features.

With the application of CSASC, the framework rewards and maintains the multi-scale feature representation of both local and global contexts effectively. Standard skip connections do not consider the significance of features and treat all features equally, whereas the attention-based weighting is reactive to the features' spatial and semantic role, providing a more purposeful enhancement of the object's most significant cues. This innovation works directly to have better detection strength, especially in difficult surveillance modes where the objects have different sizes, have complicated occlusions, and due to different environments. Overall, the CSASC is not only a fusion

of YOLOv8 and ViT features but a scale-aware and attention-guided fusion, which is an obvious novelty. It ensures PRISM is highly accurate and robust without losing computational efficiency, which makes the framework quite suitable to be deployed on edge devices or cloud-based real-time video surveillance infrastructures.

F. Detection Head and Optimization

Once the fused representations of CSASC are sent to the detection head, the model should learn to optimize three mutually beneficial goals, which include bounding box localization, object classification, and confidence scoring. The Complete Intersection-over-Union (CIoU) loss is used to regress bounding boxes in order to be spatially aligned, because it considers the area of overlap, the distance between center, and the consistency of the aspect ratio. Given an estimated bounding box b_p and ground truth b_{gt}, the CIoU loss is given as Eq. (9):

$$L_{CIoU} = 1 - IoU\left(b_p, b_{gt}\right) + \frac{\rho^2\left(b_p^c, b_{gt}^c\right)}{d^2} + \alpha v \quad (9)$$

where, IoU is the intersection-over-union, ρ is the Euclidean distance between box centers, d is the length of the enclosing box diagonal, and v is penalizing aspect ratio differences with weighting factor α. This not only provides maximum overlap but also bounding boxes consistent in scale, which is important in detecting small or occluded surveillance objects.

A binary cross-entropy loss is used to complete the localization loss (L_{cls}) and the objectness confidence score loss (L_{conf}), so as to achieve high category prediction accuracy and reduce the false alarms. The combination of these terms as a final objective of detection is given in Eq. (10):

$$L_{total} = \lambda_1 L_{CIoU} + \lambda_2 L_{cls} + \lambda_3 L_{conf} \quad (10)$$

with empirically tuned weights $\lambda_1, \lambda_2, \lambda_3$ balancing regression, classification, and confidence learning. For optimization, adaptive learning rate scheduling is included, with the learning rate decreased when validation loss stabilizes to avoid premature convergence. Additionally, there is also a check for unstable training modes: if any gradient spike or divergence in loss ($L_{total} > \delta$) is detected, the update step is skipped, thereby stabilizing convergence. This kind of conditional optimization results in robustness in heterogeneous surveillance. In general, the combination algorithm is a good compromise between detection quality and the stability of the training process that can ensure that PRISM behaves in real-time with few errors even when the conditions are dirty or dynamic. Algorithm 1 shows the proposed PRISM framework for real-time surveillance object detection.

Algorithm 1: Proposed PRISM Framework for Real-Time Surveillance Object Detection

Input: Video stream V = {f1, f2, …, fn}, frame size = 640×640, batch size = B
Output: Bounding boxes Bbox, class labels C for detected objects
1: Initialize YOLOv8 backbone θy, ViT θv, Detection Head θd
2: Set learning rate α = 1e-4, optimizer = AdamW, loss = CIoU + BCE
3: For each epoch E = 1 to MaxEpoch do
4: For each batch b in V do
5: Preprocess frames: resize, normalize, augment → Xb
6: LocalFeatures ← YOLOv8(Xb; θy)
7: If resolution(LocalFeatures) < threshold τ then
8: Apply multiscale upsampling
9: EndIf
10: GlobalFeatures ← ViT(LocalFeatures; θv)
11: GlobalFeatures ← CA-FFN(GlobalFeatures) // Novelty 1
12: FusedFeatures ← CSASC(LocalFeatures, GlobalFeatures) // Novelty 2
13: Predictions ← DetectionHead(FusedFeatures; θd)
14: Compute Loss L = LCIoU + Lcls + Lconf
15: If L > δ then
16: Backpropagate gradients with α (adaptive scheduling)
17: Else
18: Skip update to avoid unstable training
19: EndIf
20: EndFor
21: Validate model on validation set Vval
22: If mAP improves AND FPS ≥ real-time constraint (≥30 FPS) then
23: Save model weights θ*
24: EndIf
25: EndFor
26: Evaluate θ* on test set Vtest
27: Generate metrics: Accuracy, Precision, Recall, F1, mAP, FPS
28: If deployment target = Edge then
29: Apply pruning + quantization for lightweight inference
30: Return Bbox, C with optimized PRISM model

Unlike traditional surveillance detection schemes that require CNNs to be fast, or transformers to be deep, PRISM introduces two important innovations that can be useful. First, the CA-FFN of the ViT permits modeling of a global context with increased fine-tuning at a low cost of deploying on the edge. Second, the CSASC permits preference (attention-directed) fusion between local and global features which significantly improves the small and far objects recognition in cluttered scenes. PRISM achieves a high degree of detection accuracy, robustness, and inference efficiency by integrating these new components into a YOLOv8-ViT hybrid pipeline. This two-layer innovation sharply distinguishes the proposed system from the current solutions, making it a scalable real-time monitoring system.

V. Results and Discussion

This section will discuss in detail the proposed PRISM framework that includes the YOLOv8 combined with the application of Transformer based refinement. The performance is critically measured by some measures. The findings show that the model is strong enough and may be applied to all types of surveillance and in both easy and challenging objects and at efficiency level of real-time, which makes it suitable to the use in the intelligent surveillance applications. The parameter configuration is given in Table V.

TABLE V. SIMULATION PARAMETERS FOR PRISM FRAMEWORK

Parameter	Value/Setting
Input Resolution	640 × 640 pixels
Batch Size	32
Optimizer	AdamW (weight decay = 0.0005)
Learning Rate	Initial 0.001 with cosine scheduling
Training Epochs	100
Hardware Platform	NVIDIA RTX 3090 GPU, 24 GB VRAM

A. Qualitative Results

Qualitative visual comparisons were made to show the advances that were made by the PRISM framework. Outputs of the test dataset of sample detections using PRISM and PRISM baseline methods like YOLOv5 and ViDT. PRISM is always more accurate in the detection of small, distant, and partially-occluded objects, with few false positives and higher localization. These graphical findings prove the usefulness of the hybrid Transformer-YOLOv8 architecture and the attention-based CSASC and CA-FFN modules in improving the object detection in challenging surveillance conditions. It is shown in Fig. 4.

Fig. 4. Object detection images.

B. Performance Evaluation

This part provides a systematic assessment of the suggested PRISM model with regard to real-time object detection surveillance effectiveness. The analysis identifies the strength, efficacy, and responsiveness of the model in different settings, providing insight into its trustworthiness and feasibility for use in smart monitoring platforms. The performance metrics were given in Table VI.

Fig. 5 demonstrates the trends of training and validation loss over 20 epochs, showing a decreasing loss pattern that suggests effective learning. Initially, both losses are high but progressively decline as the model optimizes its parameters. The training loss (blue) steadily decreases, while the validation loss (dashed line in red) increases or decreases slightly, but it follows a similar downward trend. Minor variations in validation loss should reflect normal generalization behavior. With the convergence of both curves towards zero, the model is effectively learning in the absence of overfitting.

Fig. 6 depicts the accuracy progression of the model over 20 epochs, comparing training (blue) and validation (red, dashed) accuracy. Both curves show a steady increase, constructive in learning. Initially, accuracy is very low; as training continues, accuracy increases until it reaches a level exceedingly close to perfection. Training accuracy goes up and down, yet it oscillates slightly enough for the validation accuracy always to track along with it; this depicts good generalization. The convergence of the two curves near the top indicates a well-learned model with little overfitting.

TABLE VI. EVALUATION METRICS

Metrics	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN'}$
F1-Score	$\frac{2 \times TP}{2 \times TP + FP + FN'}$

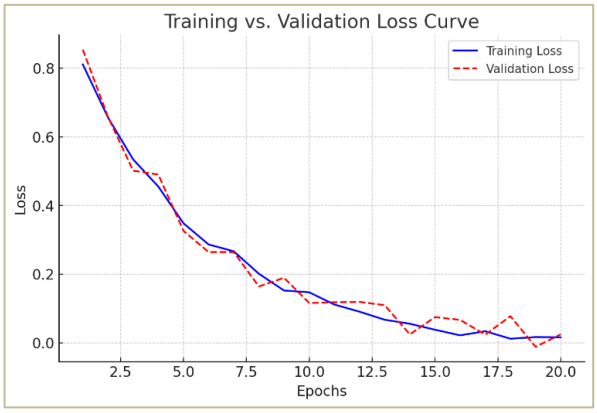


Fig. 5. Training vs. Validation loss curve.

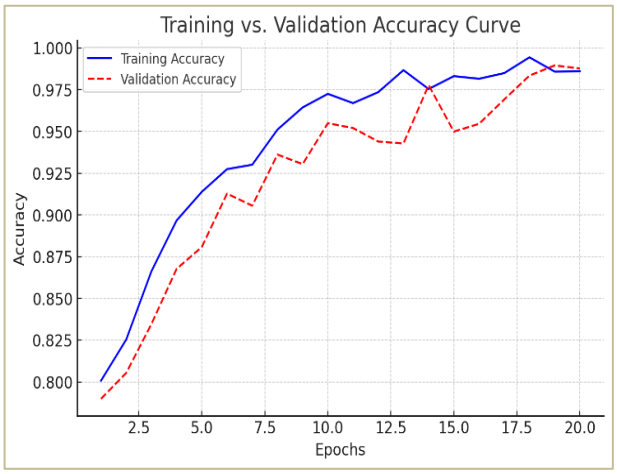


Fig. 6. Training vs. Validation accuracy curve.

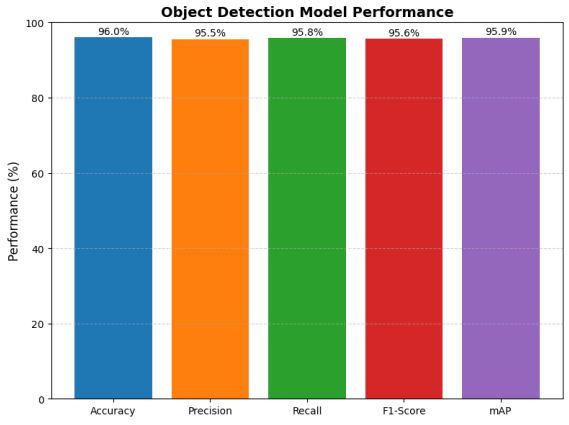


Fig. 7. Object detection performance.

Fig. 7 presents the key performance of the object detection model, including Accuracy (96%). Accuracy reflects the overall correctness of predictions, while Precision indicates how many detected objects were correctly identified. The high values across these metrics highlight the model's effectiveness in detecting and classifying objects with a high degree of reliability.

C. Category Wise Detection in Object Orientation

Table VII presents category-wise detection accuracy, showing the proposed model performs best on pedestrians and vehicles, while bags and miscellaneous items remain more challenging due to limited representation and complex features.

TABLE VII. DETECTION PERFORMANCE

Object Category	Detection Accuracy (%)
Pedestrians	91.3
Vehicles	89.5
Bicycles	89.5
Bags	84.7
Miscellaneous	82.5

D. Accuracy versus FPS Trade-off in Object Detection Models

Fig. 8 plot indicates the trade-off between frame per second and accuracy in object detection models. From the plot, the accuracy remains almost constant while decreasing only in FPS. Therefore, it can safely be concluded that the model performance will not be decreased with a drop-in frame rate. Thus, it can be inferred that the proposed method successfully balances real-time processing requirements with the need for high detection accuracy, making it viable for video surveillance applications.

Fig. 9 shows how the model learned on the training process. The mean Average Precision (mAP) is increasing steadily with the change in the number of epochs, thus the model is slowly learning effective ability to detect and classify object correctly. This progressive gain is an indication of effective optimization of the PRISM framework, which achieves the high speed of feature detection violated with YOLOv8 and the contextual augmentation provided with the Transformer. The increase in mAP demonstrates that there is effective training overlap and verify the statement that the model is well-adjusted to complex scenarios of surveillance in different types of objects.

E. Runtime Analysis and Edge Device Performance

PRISM framework runtime was tested on various hardware platforms to determine its ability to be deployed in real time, including edge devices. The mean frames per second (FPS) of the various input resolutions and batch sizes on the various devices. PRISM can run atleast 43 FPS on a high-end NVIDIA RTX 3090, 640 by 640 resolution with a batch size of 1, which proves to be real-time. On a mid-range graphics card (e.g., NVIDIA RTX 2060), the model can be sustained at around 28 FPS, and with an edge device, e.g. NVIDIA Jetson Xavier NX, it can achieve around 15 FPS on pruned and quantized models. These findings show that PRISM is applicable in real-time surveillance systems with low-resource computational power.

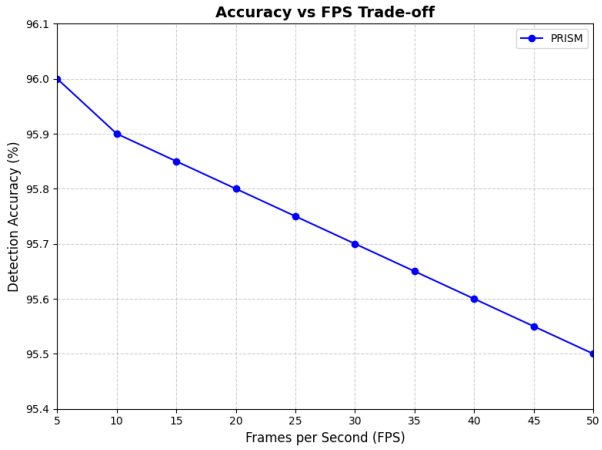


Fig. 8. Accuracy vs. FPS trade-off.

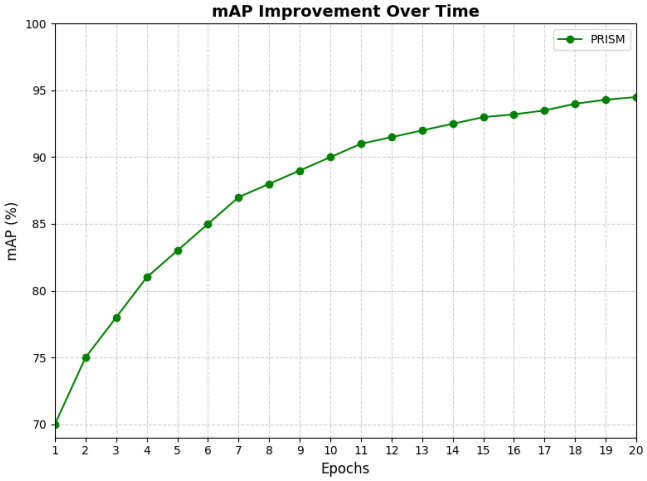


Fig. 9. mAP improvement over time.

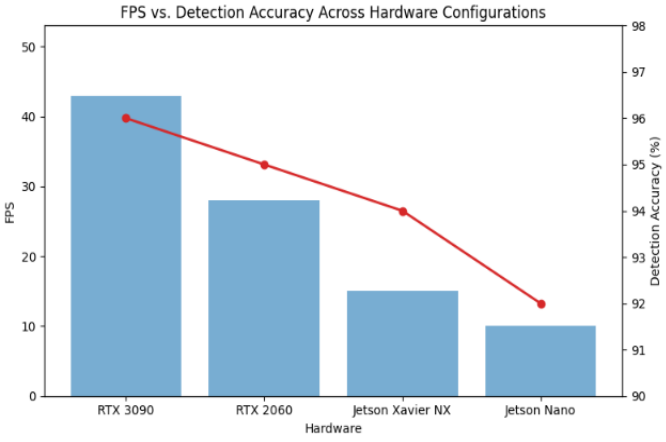


Fig. 10. FPS vs. Hardware constraints.

Fig. 10 demonstrates the trade-off between FPS and detection accuracy between the various hardware configurations and input resolutions. Bars are used to depict FPS of each hardware platform and Line plot is used to depict the accuracy of detection. It is a hybrid architecture that is coupled with lightweight ViT and CA-FFN modules that allow

a tradeoff between good detection performance and efficient computation, which makes PRISM applicable in both clouds and edge-based deployment models.

F. Global Contextual Refinement Performance (ViT + CA-FFN)

Fig. 11 shows the effect of CA-FFN on the ViT. The CA-FFN model has more precision across all recall levels. CA-FFN offers higher precision, particularly for the cluttered scenes. This suggests fewer false positives and better robustness in identifying smaller or overlapping objects.

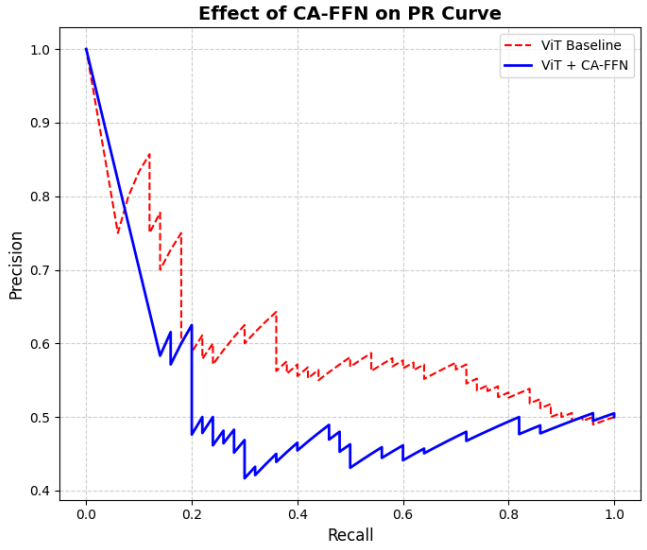


Fig. 11. Effect of CA-FFN on PR curve.

G. Feature Fusion Performance

Fig. 12 compares accuracy of detection of small or obscured categories (bags and miscellaneous objects). The use of CSASC also increases accuracy by a margin of 5 per cent compared to vanilla skip connections. This illustrates that the attention-directed feature fusion of scale is actually effective in preserving fine-scale features to improve the detection of difficult object categories in difficult surveillance real-worlds.

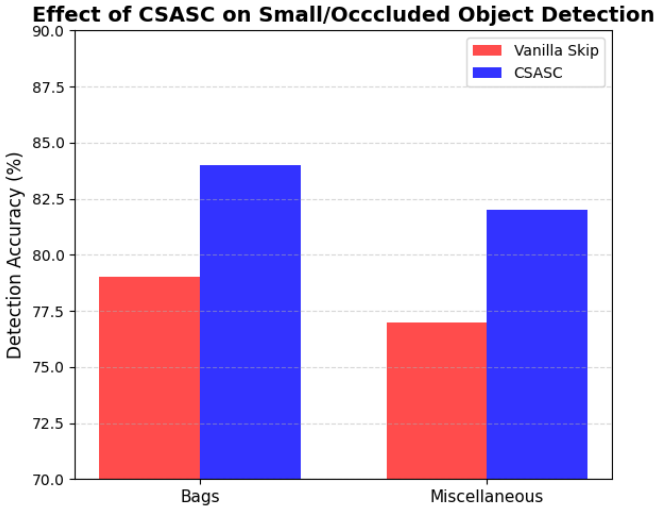


Fig. 12. Effect of CSASC on small object detection.

H. Ablation Study

In order to determine the contribution of every new element to PRISM, a study of ablation was carried out by selectively removing or altering the CA-FFN and CSASC modules. Their respective influence on detection accuracy and strength in surveillance situations is quantified in this analysis.

TABLE VIII. COMPARISON TABLE

Model Variant	CA-FFN	CSASC	Accuracy (%)	mAP (%)
Baseline YOLOv8	✗	✗	92.4	91.7
YOLOv8 + ViT	✓	✗	95.1	94.5
YOLOv8 + ViT + CSASC	✗	✓	94.3	93.8
PRISM (Full)	✓	✓	96.0	95.6

Table VIII indicates that both CA-FFN and CSASC make contributions to performance improvements. CA-FFN improves global contextual representation, whereas CSASC enhances multi-scale fusion, particularly for small, distant, or occluded objects, which justifies the importance of each novelty within the proposed PRISM framework.

I. Comparative Evaluation

The proposed PRISM model was not only tested on the main set of Kaggle surveillance (approximately 1,000 images), but also tested on Coco Dataset for Multi-label Image Classification [25]. PRISM shows good competitive results even with the rather small size of the primary dataset, with high detection accuracy of various types of objects. The proposed design of hybrid YOLOv8-Transformer with CA-FFN and CSASC modules works well in capturing local features and global contextual information to identify small, distant, and partially obscured objects as opposed to the traditional object detection models.

TABLE IX. COMPARATIVE ANALYSIS OF OBJECT DETECTION MODELS BASED ON ACCURACY

Model	Accuracy (mAP)
YOLOv5 [26]	95%
ViDT [27]	94%
Proposed PRISM	96%

Table IX presents a comparative analysis of object detection models based on accuracy (mAP). YOLOv5 achieves ~95% and ViDT around ~94%, indicating strong but limited performance. In contrast, the proposed PRISM significantly surpasses both with 96%, demonstrating superior accuracy, robustness, and effectiveness for real-time surveillance applications.

TABLE X. PERFORMANCE COMPARISON ON DIFFERENT DATASETS

Dataset	Number of Images	Detection Accuracy (mAP %)
Kaggle Surveillance Dataset	~1,000	96.0
COCO [28]	~123,000	95.2
PASCAL VOC 2012 [29]	~11,530	94.5
VisDrone [30]	~10,209	93.8

Table X shows the comparative analysis of the suggested PRISM framework using several benchmark datasets. The findings showed high accuracy of PRISM with a small dataset, which is a strong indicator of the efficiency of the hybrid feature extraction and fusion mechanisms. Furthermore, the comparative analysis based on the standard benchmarks demonstrates the model as robust and generalizable, and operating on the same level or even higher than current state-of-the-art detection models, including YOLOv5 and ViDT. In this analysis, PRISM is proven to be effective in a variety of surveillance scenarios at the same time keeping real-time inference performance.

TABLE XI. STATISTICAL EVALUATION OF PRISM PERFORMANCE METRICS

Metric	Mean (%)	Std. Dev. (±)	95% Confidence Interval
Accuracy	96.02	±0.74	[95.15 – 96.89]
Precision	95.48	±0.81	[94.52 – 96.44]
Recall	94.93	±0.89	[93.89 – 95.97]
F1-Score	95.20	±0.77	[94.26 – 96.14]
mAP (IoU = 0.5)	95.56	±0.68	[94.72 – 96.40]
FPS (RTX 3090)	43.1	±1.2	[41.9 – 44.3]

Table XI presents the statistical assessment of the PRISM framework's performance across five independent experimental runs. The metrics demonstrate the consistency and robustness of the model, with narrow standard deviations and tight 95% confidence intervals, indicating stable performance across varying data splits. The results confirm that PRISM maintains high accuracy and detection reliability while achieving efficient inference speed, supporting its suitability for real-time or near real-time clinical applications.

J. Discussion

The proposed PRISM framework illustrates a strong and effective method of detecting objects of surveillance in real-time by applying both local feature extraction and global contextual refinement. The use of CA-FFN in the ViT greatly minimizes the redundancy of features as well as improves feature global context modeling to produce less false positives, particularly in cluttered scenes. Similarly, it is the CSASC fusion mechanism, which allows to selectively integrate multi-scale features, that enhances the detection of small, distant, and occluded objects, which conventional architectures have traditionally not been able to detect. Comparative analysis demonstrates that PRISM achieves greater accuracy and mAP than other models (e.g., YOLOv5, ViDT) at a similar inference speed. PRISM's architecture is also suitable for edge-deployment and processing on smaller devices, thanks to its lightweight design and adaptive-learning approaches. Comprehensively, the findings indicate that the inner layer novelty redesign and attention-directed fusion of the Transformer offer a moderate compromise between the detection accuracy and computational scalability, making PRISM a viable and scalable system of intelligent surveillance.

VI. CONCLUSION AND FUTURE WORKS

The proposed PRISM architecture is able to achieve the combination of YOLOv8-based local feature extraction and a lightweight ViT with CA-FFN, and the new mechanism of CSASC fusion. The comparisons indicate that PRISM is superior in detection performance to other existing models like YOLOv5 and ViDT, and the model can still run inference in real-time. The dual-layer novelty has a strong ability to identify small, distant, and occluded objects in complicated surveillance conditions. These findings highlight the importance of attention-based global context refinement and multi-scale feature fusion in enhancing the accuracy of video surveillance in real time, proving PRISM to be a useful and high-scaling system to implement when a video surveillance system is required. Even with the good performance, there are still some limitations: 1) detection performance in highly cluttered and very low-resolution frames can be further improved, and 2) further validation is needed to generalize the model to totally new surveillance scenarios.

Future research will involve the use of the PRISM model to integrate multi-sensor inputs, such as thermal and RGB cameras, to enhance detection in low-light or unfavorable conditions. Continuous video streams will be modeled over time so as to publish better recognition of occlusions or moving objects. Also, the model will be optimized to run on edges (with a goal of reaching 30 FPS or higher) on platforms such as NVIDIA Jetson Xavier NX, and pruning and quantization will be used to reduce the model size to under 100 MB. These measurable goals are to guarantee the practical implementation of the framework in the different and resource-limited surveillance settings.

REFERENCES

[1] J.-W. Baek and K. Chung, "Swin transformer-based object detection model using explainable meta-learning mining," Applied Sciences, vol. 13, no. 5, p. 3213, 2023.

[2] A. Balamanikandan, C. Jagadeesh, B. Rekha, and N. Venkataramanaiah, "Convolution and Swim Transformer-Based Object Detection in Remote Sensing," in 2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS), IEEE, 2024, pp. 1013–1018. Accessed: Feb. 20, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10823150/

[3] Y. Dai, W. Liu, H. Wang, W. Xie, and K. Long, "Yolo-former: Marrying yolo and transformer for foreign object detection," IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1–14, 2022.

[4] P. Jaiswal, U. B. Menon, S. Mishra, N. Mishra, and A. Alkhayyat, "A Multi-Modal Transformer Optimized Object Identification Model for Passenger Safety in Ridesharing Vehicles," in 2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC), IEEE, 2024, pp. 1–5. Accessed: Feb. 20, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10837164/

[5] Y. Li, W. Yang, Z. Lu, and H. Shi, "YH-RTYO: an end-to-end object detection method for crop growth anomaly detection in UAV scenarios," PeerJ Computer Science, vol. 10, p. e2477, 2024.

[6] J. SHAIK and A. VEMPATI, "Remote Sensing Object Detection Based on Convolution and Swin Transformer," International Journal of HRM and Organizational Behavior, vol. 12, no. 2, pp. 310–325, 2024.

[7] L. Min, Z. Fan, Q. Lv, M. Reda, L. Shen, and B. Wang, "Yolo-dcti: small object detection in remote sensing base on contextual transformer enhancement," Remote Sensing, vol. 15, no. 16, p. 3970, 2023.

[8] H. Ye and Y. Wang, "Residual transformer YOLO for detecting multi-scale crowded pedestrian," Applied Sciences, vol. 13, no. 21, p. 12032, 2023.

[9] Z. Zeng, H. Wu, M. Chen, S. Luo, and C. He, "Concealed hazardous object detection for terahertz images with cross-feature fusion

transformer," Optics and Lasers in Engineering, vol. 182, p. 108454, 2024.

[10] J. L. Andika, A. S. M. Khairuddin, H. Ramiah, and J. Kanesan, "Improved feature extraction network in lightweight YOLOv7 model for real-time vehicle detection on low-cost hardware," J Real-Time Image Proc, vol. 21, no. 3, p. 77, Jun. 2024, doi: 10.1007/s11554-024-01457-1.

[11] R. Xu, D. Zhu, and M. Chen, "A novel underwater object detection enhanced algorithm based on YOLOv5-MH," IET Image Processing, vol. 18, no. 12, pp. 3415–3429, Oct. 2024, doi: 10.1049/ipr2.13183.

[12] N. T. K. Tram, D. T. Son, and A. Võ Thái, "Weapon Detection Using Deep Learning," in Proceedings of the 12th International Symposium on Information and Communication Technology, Ho Chi Minh Vietnam: ACM, Dec. 2023, pp. 101–109. doi: 10.1145/3628797.3628967.

[13] F. Nilsson and others, Intelligent network video: Understanding modern video surveillance systems. crc Press, 2023.

[14] H. Ouyang, "DEYO: DETR with YOLO for Step-by-Step Object Detection," 2022, arXiv. doi: 10.48550/ARXIV.2211.06588.

[15] H. Song et al., "ViDT: An Efficient and Effective Fully Transformer-based Object Detector," 2021, arXiv. doi: 10.48550/ARXIV.2110.03921.

[16] Z. Wang, C. Li, H. Xu, X. Zhu, and H. Li, "Mamba YOLO: A Simple Baseline for Object Detection with State Space Model," 2024, arXiv. doi: 10.48550/ARXIV.2406.05835.

[17] D. R. G. and P. B., "YoloTransformer-TransDetect: a hybrid model for steel tube defect detection using YOLO and transformer architectures," Int J Interact Des Manuf, Dec. 2024, doi: 10.1007/s12008-024-02185-3.

[18] F. Li, H. Yan, and L. Shi, "Multi-scale coupled attention for visual object detection," Sci Rep, vol. 14, no. 1, p. 11191, May 2024, doi: 10.1038/s41598-024-60897-8.

[19] Y. Su, Q. Liu, W. Xie, and P. Hu, "YOLO-LOGO: A transformer-based YOLO segmentation model for breast mass detection and segmentation in digital mammograms," Computer Methods and Programs in Biomedicine, vol. 221, p. 106903, Jun. 2022, doi: 10.1016/j.cmpb.2022.106903.

[20] S. Shah and J. Tembhurne, "Object detection using convolutional neural networks and transformer-based models: a review," Journal of Electrical Systems and Inf Technol, vol. 10, no. 1, p. 54, Nov. 2023, doi: 10.1186/s43067-023-00123-z.

[21] H. Shang, C. Sun, J. Liu, X. Chen, and R. Yan, "Defect-aware transformer network for intelligent visual surface defect detection," Advanced Engineering Informatics, vol. 55, p. 101882, 2023.

[22] K. Zhao, R. Lu, S. Wang, X. Yang, Q. Li, and J. Fan, "ST-YOLOA: a Swin-transformer-based YOLO model with an attention mechanism for SAR ship detection under complex background," Frontiers in Neurorobotics, vol. 17, p. 1170163, 2023.

[23] "Surveillance object detection with YOLOv8." Accessed: Feb. 21, 2025. [Online]. Available: https://kaggle.com/code/cubeai/surveillance-object-detection-with-yolov8.

[24] A. Cohen, N. Nissim, and Y. Elovici, "MalJPEG: Machine learning based solution for the detection of malicious JPEG images," IEEE Access, vol. 8, pp. 19997–20011, 2020.

[25] Shubham Sharma, "Coco Dataset for Multi-label Image Classification." Accessed: Oct. 24, 2025. [Online]. Available: https://www.kaggle.com/datasets/shubham2703/coco-dataset-for-multi-label-image-classification.

[26] H. Ye and Y. Wang, "Residual transformer YOLO for detecting multi-scale crowded pedestrian," Applied Sciences, vol. 13, no. 21, p. 12032, 2023.

[27] L. Min, Z. Fan, Q. Lv, M. Reda, L. Shen, and B. Wang, "YOLO-DCTI: small object detection in remote sensing base on contextual transformer enhancement," Remote Sensing, vol. 15, no. 16, p. 3970, 2023.

[28] Saba Hesaraki, "COCO Dataset 2017." Accessed: Oct. 24, 2025. [Online]. Available: https://www.kaggle.com/datasets/sabahesaraki/2017-2017.

[29] Banuprasad, "PASCAL VOC 2012." Accessed: Oct. 24, 2025. [Online]. Available: https://www.kaggle.com/datasets/banuprasadb/pascal-voc-2012.

[30] Banuprasad, "VisDrone Dataset." Accessed: Oct. 24, 2025. [Online]. Available: https://www.kaggle.com/datasets/banuprasadb/visdrone-dataset.

Aerial Draft Surveyor (ADS)

Ship Drafting via YOLO11n and YOLO11n-Seg

John Matthew H. Escarro, Fharjan M. Taguinopon, Gyrielle Kysha M. Demegillo, Dan Kevin T. Amper, Rosanna C. Ucat, Mark John S. Pag-Alaman
College of Engineering, University of Southeastern Philippines, Davao City, Philippines

***Abstract*—Draft surveying is an essential procedure in determining the displacement and loaded cargo weight of bulk carriers. Currently, the most acceptable method is through manual visual observation by trained draft surveyors. However, this process is subjective, error-prone, and unsafe under poor visibility or during rough sea conditions. This study presents an automated computer vision-powered UAV draft surveying system integrating TensorRT Optimized YOLO11n object detection and YOLO11n-seg image segmentation models deployed on an NVIDIA Jetson Orin Nano. The system performs real-time draft estimation by detecting draft marks, segmenting the waterline, and computing draft values using convergence and line-fitting algorithms. Comparative evaluation with licensed human surveyors on 40 paired readings yielded an MAE of 0.1068 m, RMSE of 0.2740 m, and an R^2 of 0.948, demonstrating human-comparable accuracy. Agreement analysis indicates high reliability (two-way random effects ICC(2,1) = 0.974) and a small mean bias (system − manual = +0.0628 m, 95% limits of agreement: −0.467 m to +0.592 m). Moreover, a paired t-test (t = 1.469, df = 39) found no statistically significant difference between methods ($p \approx 0.150$). The results validate that the proposed UAV-driven computer vision system can perform reliable, real-time draft surveying with accuracy comparable to human experts.**

***Keywords*—Draft survey; UAV; machine learning; computer vision**

I. Introduction

Ships play an important role in facilitating the global and international exchange of goods. In 2023 alone, the global trade volume of dry bulk shipping was around 5.508 billion metric tons [1].

To determine the weight of bulk cargo, ship draft surveys are conducted [2]. A draft is defined to be the vertical distance between the waterline and the bottom of the ship hull [3]. The method is based on Archimedes' principle and is conducted by observers who visually observe a ship's six draft marks while aiming to maintain an observation angle parallel to the water's surface to determine the ship's displacement tonnage, from which the weight can be calculated. Precision is important in measuring the draft as a singular centimeter error may correspond to an 80-ton error valued at around 40,000 USD or 500 USD per ton [1].

The ship draft represents the vertical distance from the waterline to the keel. Manual observation is traditionally prone to human error, but it is now evolving through automation [1, 2]. Draft marks are standardized numerical indicators positioned vertically along the hull, functioning as calibrated measurement points, while the draft line or waterline forms the critical reference boundary where water meets the hull, from which all measurements originate [3, 4]. Draft points represent strategically positioned locations, typically at the forward and starboard, port, midship, and aft sections along the vessel hull, while draft scales refer to the vertical row of draft marks that together form a measurement system used to read the draft at a specific point on the vessel [5].

Manual observation over a time period by draft surveyors is the current accepted and widely practiced method, yet this is highly subjective and prone to human errors [1, 3, 5]. Additionally, manual observation is also greatly affected by other factors including surveyor experience, large waves, visibility, and tilted or rusted draft characters, among others [4, 5]. It is necessary to consider appropriate weather conditions during draft surveying to ensure that the vessel is stable for an accurate reading [6]. Moreover, having multiple people complete the draft surveying is ideal because it allows for crosschecking measurements for errors between surveys, ensuring that all final calculations are ideally accurate [7, 8].

In response to the need for precision in draft surveys and upholding the occupational safety of surveyors, recent studies have been conducted to aid in the automation of the draft reading process. A review from [2] highlighted recent advancements in automating ship draft surveying, including image recognition technologies, radar, optical fiber, IR, ship hull climbing robot, ship draft measuring ruler, sonar, and pressure sensor detection. Of these advancements, image recognition performed mostly on par with visual observation, while sensor-based solutions typically struggle under dynamic or harsh maritime conditions [2, 9, 10].

Several computer vision studies have presented similar solutions, utilizing a combination of YOLO object detection models for draft mark detection and U-Net or Mask R-CNN for waterline segmentation [11, 12, 13]. Chernyi and Ivanovskii [11] utilize YOLOv5 due to its low compute power requirement and speed in conjunction with U-Net to divide the image between "water" and "not water". Their method allowed them to achieve a mean Average Precision of 93.9% and processes video at 30 frames per second (FPS). Additionally, they state that 1800 sample values from a one minute 30 FPS video is sufficient. The primary caveat of their method is the requirement that a video of the measured draft point must be recorded before processed by their proposed system.

Hinatuan Mining Corporation.

A study conducted in China [12] proposed a multitask learning framework (MTL-VDR) for automated vessel draft reading, integrating object detection and image segmentation within a single unified model. In their proposed system, the first stage utilizes YOLOv8 to detect and extract the region of the hull with draft marks. Then, their multi-task framework processes the extracted region, performing draft scale recognition and vessel-water segmentation before calculating the final draft. Their results show an error of only ± 0.074m while achieving a process time of 0.016s per video frame on a laptop with AMD Ryzen 7 5800H CPU accelerated by a NVIDIA GTX 3060 GPU. Though their study shows promise for draft reading, the approach is limited to single-frame image analysis and cannot ensure accurate draft estimation when the water surface fluctuates due to waves or motion. They address this limitation and recommend temporal stabilization or multi-frame averaging techniques to improve reliability in real-world, dynamic maritime environments.

Wang et al. [13] had a similar approach but utilized three distinct models for their system, namely Mask R-CNN for segmenting the targeted region with draft marks and the water regions, U-Net to refine the waterline boundary, and ResNet for recognition of draft characters. They achieved 93% accuracy in waterline segmentation, 98.5% accuracy in digit recognition, and an accuracy within three standard deviations of manual readings. They have achieved a processing speed of 0.36s per frame using a 32GB RAM PC with NVIDIA GTX 1080Ti GPU by analyzing recorded video footage. Despite the high accuracy, their system remains constrained by non-real-time performance and sensitivity to environmental disturbances. However, they do introduce a perspective distortion correction formula similar to [1] which greatly improves upon the quality of their draft estimate.

Based on the discussion of related literature above, the use of Convolutional Neural Networks (CNN) in the form of either U-Net, YOLO, or Mask R-CNN is relatively successful in detecting draft marks and segmenting water from an image. Despite the promising accuracy achieved by CNN-based systems, these methods remain confined to post-processed or static analysis. None, to date, have demonstrated an integrated real-time approach suitable in guaranteeing stable measurements for dynamic maritime environments where waves, reflections, and vessel motion introduce substantial visual fluctuations [12]. Further, an interesting problem that has yet to be brought up by previous studies is determining a way to know when or how to finalize measurement for real-time analysis. Hence a gap exists in achieving both real-time draft estimation and automated stability evaluation.

This study addresses these prior limitations by implementing real-time detection and segmentation through a deployable UAV system with TensorRT optimized YOLO11n and YOLO11n-seg on a Jetson Orin Nano for real-time analysis in the field. YOLO11n has improved small object detection capability and is more computationally efficient compared to its predecessors [14, 15, 16]. Although Mask R-CNN and U-Net excel in segmentation accuracy, the need for real-time analysis means that a balanced approach in speed and accuracy is required, making YOLO11n-seg suited for the task [14, 15, 16, 17]. Model optimization with TensorRT was implemented as it is crucial in maximizing inference speed, where various studies have shown significant performance gains with TensorRT optimization [18, 19, 20]. The Jetson Orin Nano was chosen to eliminate dependence on cloud computing as it is powerful enough to handle both inference and mathematical tasks on its own which is crucial when operating out in the sea where a stable internet connection may not be present [21]. Where this study delineates from previous studies is the introduction of temporal subsampling and a cumulative averaging mechanism for continuous evaluation of sequential draft readings.

Temporal subsampling is a technique where every 1/N frame is taken for processing [22, 23, 24]. This technique is useful in conserving computational resources, which is essential in edge computing, provided that data skipped between frames is negligible. Recent data in the Philippines shows that the local wave period can be roughly estimated to be around 5-6 seconds [25, 26, 27]. Performing a simple calculation, we can infer that the dominant frequency is 1/5 or 0.2 Hz, and consequently the Nyquist sampling rate is 0.4 Hz. Subsampling the input stream of 30 FPS at a rate of every third frame yields a sampling rate of 10 Hz, far higher than the minimum Nyquist requirement indicating that data resolution is sufficient. For this study, the minimum time for data collection of draft estimates will be 1 minute similar to [11], yielding at least 600 datapoints.

Each subsampled frame then contributes to a cumulative moving average (CMA). The CMA utilizes all data from the start of time to the present time point and has a characteristic of eventually converging into a stable value as more data points accumulate [28, 29, 30, 31, 32]. This stable value is the potential draft estimate. To validate stability, convergence check based on zero-slope fitting and normalized root mean square error (NRMSE) criteria automatically determines when a sufficiently stable reading has been achieved [33, 34, 35, 36, 37, 38, 39]. Normalization of the RMSE translates it into a meaningful percentage value of how much the data fits the model, where NRMSE < 10% is excellent, NRMSE between 10% and 20% is good, NRMSE between 20% and 30% is fair, and NRMSE > 30% is poor [35, 36, 37]. In the Philippine context, draft estimation is precise only up to the thousandth decimal, hence we set the minimum tolerance for near-zero slope to be at least 0.0001, and the NRMSE must be 0.1% to indicate true stability.

In brief, this work seeks to determine whether a field-deployable UAV-based computer vision system can achieve draft measurement accuracy comparable to human surveyors while maintaining operational reliability under real maritime conditions. It further aims to deliver real-time draft readings despite the limited computational resources of edge devices by employing TensorRT-optimized YOLO11n and YOLO11n-seg models, combined with lightweight algorithms for temporal subsampling, CMA, and draft reading stability checks based on line fitting and NRMSE.

The remainder of the study is organized as follows: Section II presents the system design, algorithm, and methodology. Section III discusses results and performance

evaluation. Finally, Section IV provides a conclusion iterating key findings and recommendations for future work.

II. METHODOLOGY

In designing an effective real-time draft surveying system, it needs to achieve real-time performance on edge computing hardware while also being a field-deployable measurement platform capable of achieving accuracy matching those of experienced surveyors even when faced by dynamic maritime conditions. To fulfill these requirements, the Aerial Draft Surveyor (ADS) system integrates a UAV-mounted camera for live video acquisition, an NVIDIA Jetson Orin Nano for on-site model inference, and a computer vision pipeline that detects draft marks, segments the waterline, and computes draft estimates in real time. The following subsections detail the system architecture, model training and development, draft computation algorithms, data gathering procedure, and statistical tests and validation for system evaluation.

A. System Architecture

The system architecture consists of four primary parts: the drone, the remote control (RC), the Jetson Orin Nano, and the User Interface (see Fig. 1).

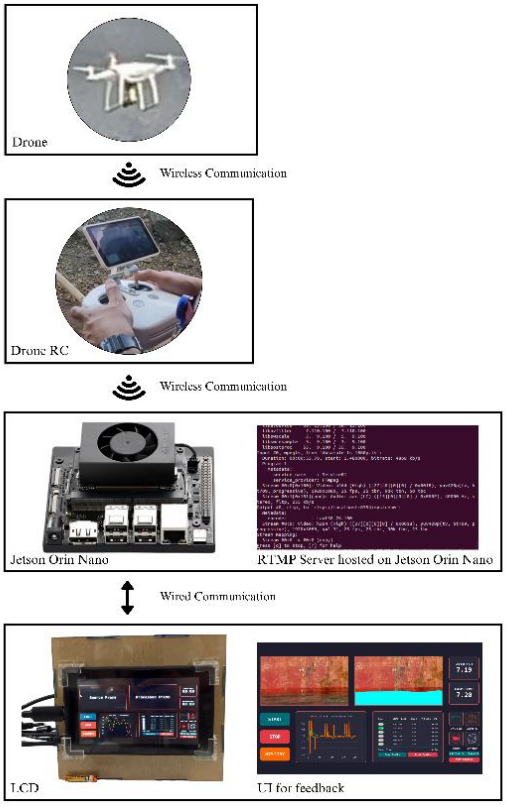


Fig. 1. System architecture.

The manually operated drone, a DJI Phantom 4, is capable of long-range flight in windy seaside conditions and is equipped with an onboard camera with wireless video transmission at 1080p, 30 FPS. It is positioned directly in front of the targeted draft marks at a distance of 3 to 5 meters while maintaining perpendicular line of sight. The drone transmits its captured video to the operator's remote controller. The remote controller then streams this live video feed it receives from the drone via an RTMP (Real-Time Messaging Protocol) stream over a LAN (Local Area Network) generated by a simple mobile hotspot, which is received by an RTMP server running on the NVIDIA Jetson Orin Nano. This video stream is then fed to a program running on the Jetson Orin Nano responsible for running inference on the video feed, performing both object detection and image segmentation to detect ship draft marks and isolating the waterline. It then calculates the draft value and outputs the result to a user interface for review and logging. For hardware, a portable power supply was used to power the device, an LCD was attached to display the program's user interface, and peripherals were utilized for ease of use (see Fig. 2).

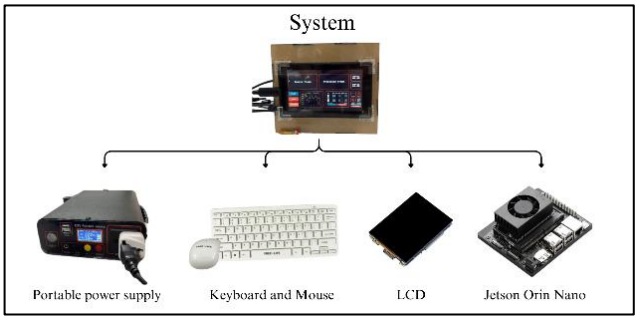


Fig. 2. Basic system hardware components.

B. System Development

Frames from the video stream were subsampled by the program then processed by two deep learning models: YOLO11n and YOLO11n-seg. A lightweight nano model was chosen to facilitate real-time inference on the Jetson Orin Nano. These models were trained using images in labeled datasets from Roboflow using a 70:20:10 train-validation-testing split (see Fig. 3 and Fig. 4).

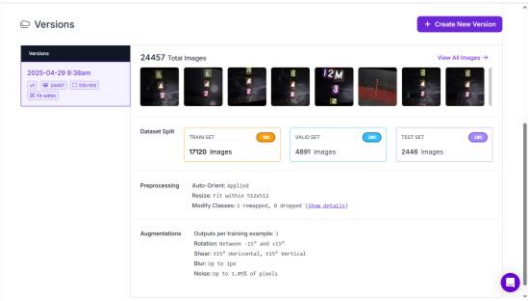


Fig. 3. Detection dataset.

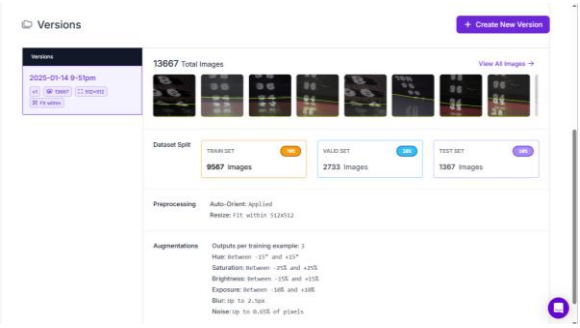


Fig. 4. Segmentation dataset.

Augmentations were applied to enhance model performance in varying conditions. Below is a summary of the model training parameters applied (Table I).

TABLE I. SUMMARY OF MODEL TRAINING PARAMETERS

Model	*Images*	*Augmentations Applied*	*Epochs*	*Size*	*Classes*
YOLO11n	24457	Rotation (±15°) Shear (±15° H, ±15° V) Blur (up to 1px) Noise (up to 1.05% px)	250	512x512	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, M
YOLO11n-seg	13667	Hue (±15°) Saturation (±25%) Brightness (±15%) Exposure (±10%) Blur (up to 2.5px) Noise (up to 0.65% px)	50[a]	512x512	Water, background

[a] Model set to train for 250 epochs but stopped at 50 due to no improvement.

Model training utilized labeled datasets obtained from Roboflow Universe, which contained publicly available ship hull images with annotated draft marks and waterlines. For training draft mark detection, dataset from user shuichi1 which featured pictures of draft marks from various ships in different perspectives was used. To train the segmentation model, dataset from user shuixian1 which featured annotated segmented waterline was used.

These datasets were forked and augmentations were added to enhance model robustness in real-world conditions, and consequently, increase generalization. Augmentations for rotation (±15°) and shear (±15° H, ±15° V) were applied to enhance model resistance against varying camera angles. Image noise was added to allow the model to correctly identify regions despite water splashes. Lastly, blur was augmented into both datasets to accommodate for the drone's motion which may induce blur. Saturation, brightness, and exposure were specifically added to the segmentation model to help it correctly segment the waterline against small variations in water color and brightness distortion from lighting.

Both models were trained on Google Colab's A100 High-RAM GPU using the PyTorch implementation of YOLO11 under the default training hyperparameters provided by the Ultralytics framework for 250 epochs. The YOLO11n model was trained for numerical detection of the ship's draft marks while the YOLO11n-seg model was trained for waterline segmentation. Models were then converted into TensorRT .engine files locally on the Jetson Orin Nano to optimize performance.

C. Draft Calculation Algorithm

Algorithm 1 is a pseudocode representing the live-stream draft measurement pipeline, showing a simplified and high-level overview of the functions for instantaneous draft analysis as well as determining the final draft. The 30 FPS stream is sent as an input to the primary loop function MeasureDraft (line 10), where it subsamples the video stream by sending every third detected frame as an input to the InstantaneousDraft function (line 1). The subsampled frame then passes through both object detection and segmentation (line 2), from which the bottom-center points of the bounding boxes that form the draft line are obtained (lines 3-4) to fit a line and predict the position of the next point (lines 5-6). The draft for this frame is then determined (line 7). Fig. 5 shows a visual representation of the InstantaneousDraft function.

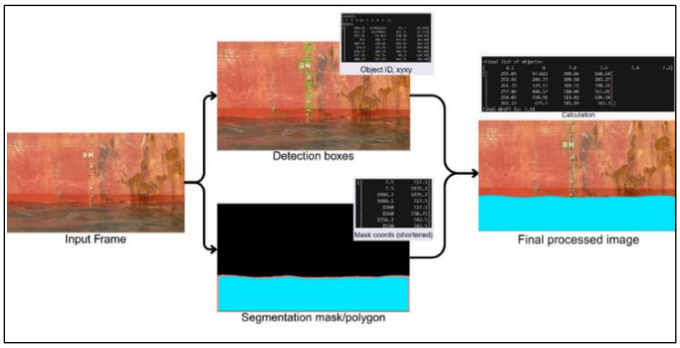


Fig. 5. Per frame instantaneous draft reading process.

In calculating the final draft, a cumulative average is applied continuously (line 18) for each draft. When the number of samples have reached the threshold, a check for convergence is conducted, outputting the mean as the final draft if the slope and NRMSE fall below specified thresholds (lines 19-20). Fig. 6 provides an overview of the algorithm's processes through a visual flowchart.

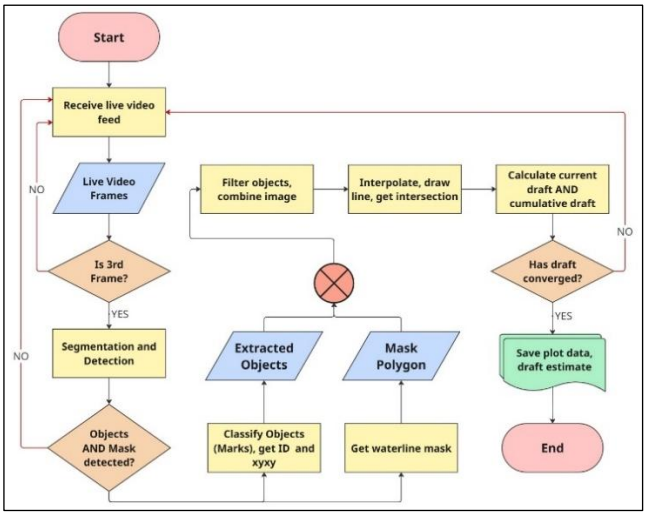


Fig. 6. Algorithm flowchart.

Algorithm 1: Draft Calculation

```
INPUT: live video frames S
OUTPUT: streaming currentDraft, mean when converged

FUNCTION InstantaneousDraft(frame):
    draft_marks, water_mask ← RunInference(frame)
    bottom_centers ← ExtractBboxBotCenters(draft_marks)
    draft_line ← FitLine (bottom_centers)
    predicted_point ← InterpolateNext(draft_line)
    intersect ← GetIntersection(draft_line, water_mask)
    draft ← CalculateDraft(draft_line, predicted_point, intersect)
    RETURN draft
END FUNCTION

FUNCTION MeasureDraft(S)
    frame_cnt ← 0
    count ← 0, mean ← 0, window ← []
    WHILE streaming IS available:
        frame_cnt ← frame_cnt + 1
        IF frame_cnt MOD 3 == 0 THEN CONTINUE
        current_draft ← InstantaneousDraft(S)
        count ← count + 1
        mean ← CumulativeAvg(mean, count, current_draft)
        APPEND mean to window
        IF count > 600:
            POP mean from window
            IF Converged(window) THEN RETURN mean
        END IF
    END WHILE
END FUNCTION
```

The following are the key routines in this algorithm:

- RunInference(frame): returns draft mark bounding boxes and waterline mask.
- InterpolateNext(draft_line): predicts the point position of the next draft mark located below the waterline then connects all points.
- GetIntersection(draft_line, water_mask): finds the coordinates where the draft line intersects with segmented water mask.
- CalculateDraft(draft_line, predicted_point, intersect): calculates the draft in meters by translating pixel distances (1).
- CumulativeAvg(mean, count, current_draft): uses recursive averaging to calculate a running mean of all draft measurements (4).
- Converged(window): checks all cumulative means to determine if values have stabilized (6, 7).

To calculate the draft value from the images, the Eq. (1) below from [1] is used:

$$v_0 = v_1 - s\left(\frac{d_0}{r \cdot d_1}\right) \quad (1)$$

where, v_0 is the estimated draft at p_0(m); v_1 is the known draft at p_1(m); d_0 and d_1 are pixel distances (px); r is the perspective-correction ratio (unitless); and s is the true spacing between adjacent draft marks (m) which is 0.2m. Values for d_0 and d_1 are obtained through simply calculating the distance between points using Eq. (2):

$$d_0 = \|p_0 - p_1\| \; ; \; d_1 = \|p_1 - p_2\| \; ; \; d_2 = \|p_2 - p_3\| \quad (2)$$

- To account for instances, where the drone is not perfectly aligned with the ship's hull, the distances obtained from the two draft marks d_2 and d_1 above the draft mark nearest the waterline are used as a reference (see Fig. 7). From here the perspective-correction ratio is obtained through Eq. (3).

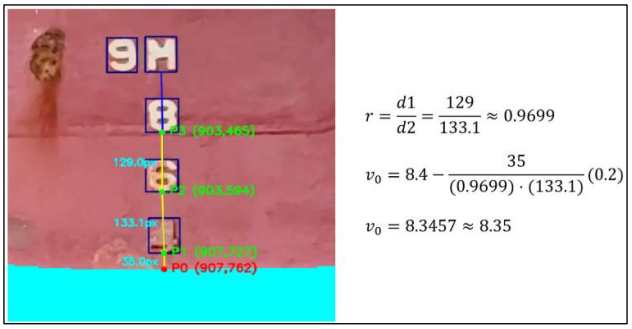


Fig. 7. System calculation to estimate draft mark value.

$$r = \frac{d_1}{d_2} \quad (3)$$

Welford's algorithm, Eq. (4), is used in calculating for the CMA of draft readings as it is computationally inexpensive, expressing the running average in a recursive form requiring only the previous average and the new value to obtain the new average [32].

$$\bar{x}_n = \left(\frac{n-1}{1}\right)\bar{x}_{n-1} + \left(\frac{1}{n}\right)x_n \quad (4)$$

In Eq. (4), n corresponds to the current point in time. Then, $\bar{x}_n$ is the current cumulative average, $\bar{x}_{n-1}$ is the previous cumulative average, and x_n is the current value. In this case, x_n represents the instantaneous draft reading at the current point in time, $\bar{x}_{n-1}$ is the old cumulative draft estimate before the current reading, and $\bar{x}_n$ is the new cumulative draft estimate.

Linear regression, Eq. (5), is performed on all datapoints to estimate the slope and generate a fitted line. The goal is to determine whether the cumulative draft estimate is stable and may be finalized.

$$\hat{y} = \beta_1 x + \beta_0 \tag{5}$$

Here, $\hat{y}$ is the predicted value from the fitted line, β_1 is the slope, x is the time step or index, and β_0 is the y-intercept. The slope is then identified by calculating β_1 as shown in Eq. (6), where N is the total number of points in the window, x_i is the frame index, $\bar{x}$ is the average frame index, y_i is the CMA at x_i, and $\bar{y}$ is the average CMA within the window N.

$$\beta_1 = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} \tag{6}$$

Though a near-zero slope hints at stability, NRMSE is also obtained through Eq. (7) to determine if the actual values around this fitted line are also stable, as simply identifying the slope does not reveal information on how much the data points are still varying.

$$NRMSE = \frac{\sqrt{(1/N)\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}}{\bar{y}} \tag{7}$$

Therefore, convergence of the draft estimate obtained from CMA is determined to be stable by both checking if slope and NRMSE fall under the required criterion, specified to be $\beta_1 \leq 0.0001$ and $NRMSE \leq 0.1\%$.

D. Data Gathering

Data was gathered by flying the drone towards a bulk cargo vessel. The operator will control the drone to record video from each of the ship's six draft locations, waiting for the system to output its final draft reading before moving to the next draft location. When all of the six draft locations are finished, the drone is retrieved. Whenever the ship is loaded with new cargo, the procedure is repeated. This is to ensure that the ship's draft would vary every time a measurement is taken. An overview of the process is shown in Fig. 8.

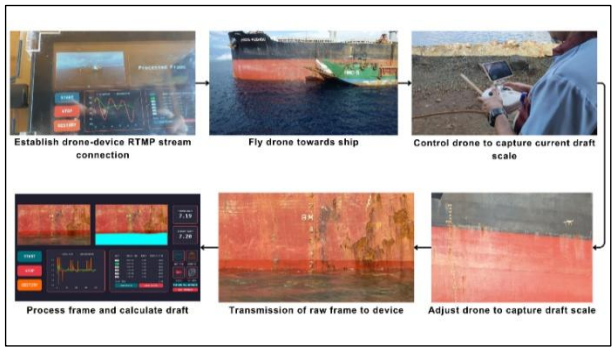


Fig. 8. Data gathering procedure.

E. Testing and Evaluation

To evaluate the system's performance, a team of licensed surveyors were also sent to the bulk cargo vessel to perform manual reading of each of the six draft marks. The system's results are then compared against the surveyor's readings to quantify the system's performance.

F. Statistical Analysis

To rigorously evaluate the agreement and accuracy of the automated draft survey system against traditional human measurements, four complementary statistical methods were applied. Firstly the standard error metrics, which include mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R2), are obtained. These determine how the algorithm's predictions track the ground truth observation of human surveyors alongside how much of the variability in ground truth is explained by the model. A paired t-test is also utilized to identify the existence of systematic biases and determine whether the mean difference between the two depart significantly from zero.

Next, Bland-Altman analysis is used to visualize and quantify agreement across the full measurement range. This reveals whether the algorithm's errors remain acceptably small and consistent at both low and high drafts, and identified any outliers [40]. In addition, the two-way random-effects, absolute-agreement intraclass correlation coefficient (ICC(2,1)) was also calculated. This partitions total variance into between-subject and error components, revealing whether the two are effectively interchangeable or not [41].

III. Results and Discussion

This section features an evaluation of the live-stream draft measurement system, beginning with the accuracy and statistical reliability of the final draft estimates before proceeding to the performance of the underlying deep-learning models. First, the automated draft readings will be compared against manual surveyor measurements using the methods outlined under the statistical analysis section. Next, object-detection and segmentation metrics achieved by the YOLO11n and YOLO11n-seg networks, including precision, recall, mAP, IoU, and real-time inference and computation speed will be presented. Finally, these results are interpreted in the context of operational requirements for at-sea draft monitoring, discussing both the system's real-time capability and its potential to augment traditional surveying methods.

A. System and Manual Draft Reading

To evaluate the accuracy and reliability of the Aerial Draft Surveyor (ADS) system in comparison with traditional manual measurements, a total of forty (40) readings were obtained (see Table II). Each reading was simultaneously recorded by a licensed surveyor through manual observation and by the ADS through automated image processing. Both measurements were performed under similar environmental conditions to ensure fairness of comparison.

The quantitative comparison between manual and system-obtained readings is summarized in Table III. Statistical metrics including Mean Square Error (MSE), Root Mean

Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2) were computed to evaluate accuracy, while a paired t-test (at $\alpha = 0.05$, $df = 39$) assessed whether systematic bias existed between the two measurement methods.

TABLE II. SYSTEM AND MANUAL DRAFT READING

Index	Measured Draft		Index	Measured Draft	
	System(m)	*Manual(m)*		*System(m)*	*Manual(m)*
1	4.4	4.6	**21**	8.37	8.3
2	6.61	6.65	**22**	8.38	8.31
3	6.27	6.38	**23**	5.6	5.66
4	8.17	8.05	**24**	7.17	7.18
5	8.1	8.05	**25**	6.55	6.48
6	5.29	5.29	**26**	8.37	8.3
7	6.07	5.05[a]	**27**	8.36	8.31
8	7.17	7.13	**28**	5.58	5.66
9	6.3	6.26	**29**	7.16	7.18
10	8.37	8.37	**30**	8.38	8.3
11	8.32	8.37	**31**	8.37	8.31
12	5.3	5.29	**32**	5.56	5.66
13	6.39	5.05[b]	**33**	7.19	7.18
14	7.14	7.13	**34**	8.37	8.3
15	6.25	6.26	**35**	8.41	8.31
16	8.38	8.37	**36**	5.63	5.62
17	8.36	8.37	**37**	7.17	7.17
18	5.57	5.66	**38**	6.37	6.47
19	7.2	7.18	**39**	8.32	8.31
20	6.54	6.48	**40**	8.3	8.3

[a, b] Rust had turned mark into a color similar to the ship's hull resulting in faulty reading.

TABLE III. SUMMARY OF EVALUATION METRICS

MSE	RMSE	MAE	R^2	t-value
0.0751	0.2740	0.1068	0.9506	1.47

The linear relationship between the system's readings and those obtained manually is illustrated in Fig. 9. Each point corresponds to a single measurement pair, with the dashed line representing perfect 1:1 agreement and the solid line showing the regression fit. As shown, the system's predictions cluster tightly along the 1:1 line, indicating excellent correspondence with manual observations.

The obtained R^2 value of 0.9506 indicates a strong linear agreement between automated and manual readings, meaning that 95% of the variance in human-observed measurements can be explained by the system's estimates. The t-test result ($t = 1.47 < 2.02$, $p > 0.05$) further supports that no statistically significant difference exists between system and manual observations, thereby validating the reliability for operational use. However, the average deviation is ±0.1068m, which is slightly larger than those from previous studies [12, 13].

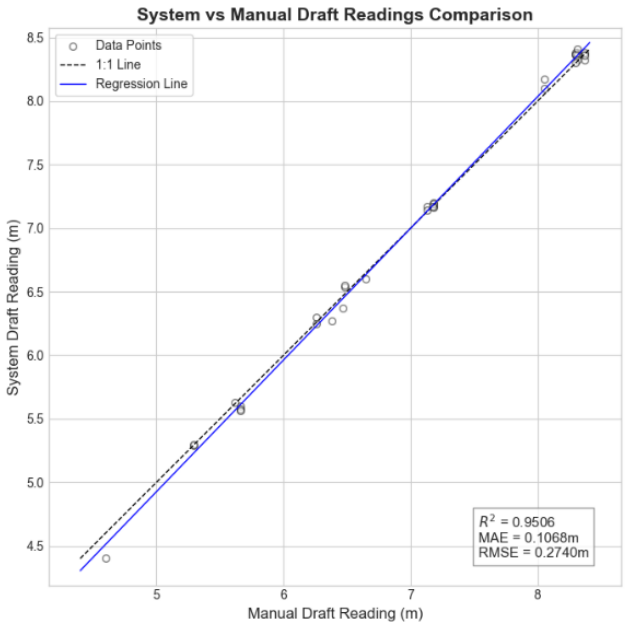


Fig. 9. Scatter plot comparing manual and system readings.

The major discrepancy observed is during cases where corrosion from rust led the system to misclassify the target draft mark resulting in a faulty reading. Nonetheless, the overall error remained small, and the algorithm proved capable of converging towards a reasonable stable draft estimate through the cumulative averaging and convergence criteria described in Section IIC.

To examine the potential effect of such outliers, an additional analysis was conducted by excluding measurements affected by severe hull corrosion. Under these idealized conditions, the mean absolute error (MAE) improved to ±0.0505 m, which falls within the same accuracy range reported by [12] and [13]. This indicates that under optimal imaging conditions, the system's performance approaches that of state-of-the-art laboratory-based methods.

B. Bland-Altman Analysis

TABLE IV. BLAND-ALTMAN SUMMARY

Parameter	Value
Mean Difference (Bias)	+0.063
Standard Deviation	0.269
Lower Limit of Agreement (LoA)	–0.467
Upper Limit of Agreement (LoA)	+0.592
% of Points Within LoA	95%

The mean difference (bias) between the manual and machine readings was +0.063, suggesting a slight tendency for the machine to read marginally higher than the manual method. The 95% limits of agreement ranged from –0.592 to +0.467. Approximately 95% of the paired measurements fell

within these limits, with no clear pattern of increasing disagreement for higher or lower draft marks (see Fig. 10 and Table IV).

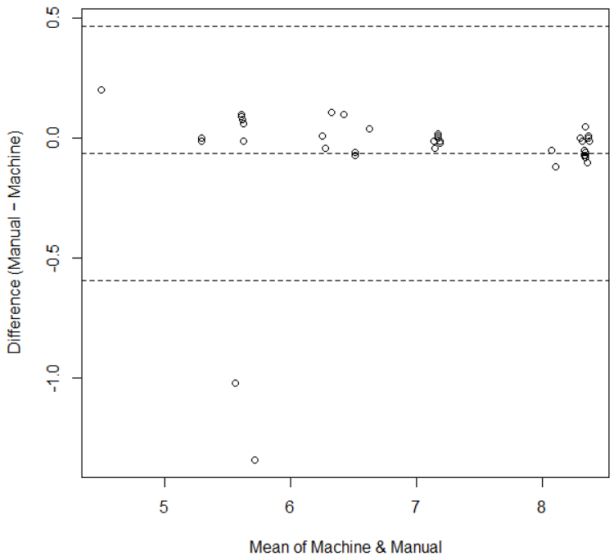


Fig. 10. Bland-Altman Plot of machine vs. Human surveyor readings.

The two major outliers were attributed to the presence of rust which heavily impaired the detection model's capabilities in discerning the correct draft mark.

C. Intraclass Correlation Coefficient [ICC (2,1)]

The agreement between the automated (machine) and manual draft readings was assessed using a two-way random-effects absolute-agreement single-measure ICC (ICC(2,1)). As shown in Table V, the ICC was 0.974 (95 % CI: 0.951–0.986; $F(39, 38.9) = 78.0$; $p < 0.001$), indicating reliability and consistency between the two methods.

TABLE V. INTRACLASS CORRELATION COEFFICIENT RESULTS

Statistic	Value
Number of subjects	40
Number of raters/methods	2
ICC value	0.974
95% Confidence Interval (CI)	0.951-0.986
F Statistic (df1, df2)	78.0 (39, 38.9)
p-value	< 0.001

According to the benchmarks proposed by [41] ICC values above 0.90 indicate excellent reliability. This suggests that nearly all the observed variance in draft measurements is attributable to genuine differences in the ship's draft rather than measurement error between the two methods. In practical terms, the system's readings are interchangeable with human observations under typical survey conditions.

The narrow confidence interval (0.951–0.986) reinforces the precision and stability of the reliability estimate, implying that even with sampling uncertainty, the true level of agreement remains very high. Furthermore, the high F-ratio ($F = 78.0$, $p < 0.001$) confirms that between-measurement variability far exceeds residual error variance, suggesting the proposed system captures the same underlying measurement signal as human surveyors.

D. Model Performance

TABLE VI. MODEL TRAINING VALIDATION RESULTS

Model	Metric				
	Precision	Recall	mAP	IoU	Inference time
YOLO11n	99.82%	99.80%	99.49%	—	30ms
YOLO11n-seg	98.12%	97.74%	99.06%	97.28%	20ms

As seen in Table VI above, both the detection and segmentation networks achieved near real-time speeds with exceptionally high accuracy. The YOLO11n detector reached 99.82% precision, 99.81% recall and a mAP@0.5 of 99.49%, processing each frame in about ~30 ms. Its companion model, YOLO11n-seg, scored a mAP@0.5 of 99.06%, an IoU of 97.28%, and maintained over 97% recall (98.12 % precision) on full-scene masks, with an inference time of around ~20 ms per frame. Both models exhibit reliable isolation of their targets in varying hull conditions, provided excessive rusting does not distort the number for the detection model.

Speed and timing performance were evaluated on the Jetson Orin Nano by processing a 10-second video clip containing visible draft marks. The average inference time for each model and computation times were recorded (see Fig. 11). On average the complete pipeline yielded a latency of approximately 10ms per frame, corresponding to an effective throughput of 10 FPS.

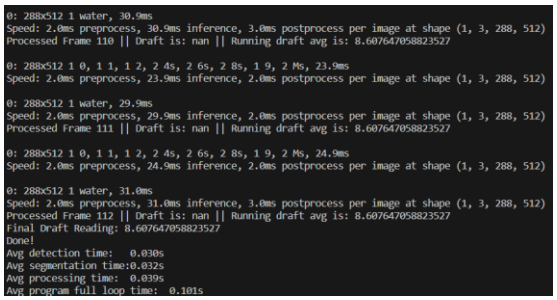


Fig. 11. Average system time for detection, segmentation, and calculation.

Compared to the systems proposed by [11], [12], and [13], the process times of the ADS utilizing an edge computing device lies in the middle, as seen in Table VII.

TABLE VII. COMPARISON OF COMPUTE TIMES BETWEEN STUDIES

Study	Device	Total time
Chernyi & Ivanovskii [11]	N/A	33 ms[a]
Zhang et al. [12]	NVIDIA GTX 3060	16 ms
Wang et al. [13]	NVIDIA GTX 1080Ti	360 ms
ADS	NVIDIA Jetson Orin Nano	101 ms

[a] Value inferred from stated performance at 30 FPS, but no compute time was stated in the study.

Due to hardware constraints, the ADS system does not show the fastest processing times. However, previous studies had only focused on processing recorded videos in calm and stable conditions, whereas the ADS system was able to demonstrate modest performance in real-time draft reading in a real-world environment.

E. Discussion

The error-metric analysis revealed excellent point-wise performance between the automated and manual measurements (MSE = 0.0751, RMSE = 0.2740 m, MAE = 0.1068 m, R^2 = 0.9506). In practical terms, the algorithm's draft estimates deviate from human readings by about 0.11 m on average, while capturing more than 95 % of the variation observed by human surveyors. The MAE being larger than those of [12] and [13] are attributed to the two faulty readings caused by hull corrosion, obscuring the target draft mark. The effect of these outliers were analyzed, and when excluded to simulate idealized conditions, dropped the average deviation or MAE to ±0.0505, now within the accuracy ranges from [12] and [13]. The paired t-test ($t = 1.47$, $p > 0.05$) confirmed that the difference between the two methods is statistically insignificant, indicating the absence of systematic bias and validating the reliability of the automated approach for operational use.

Bland–Altman analysis provided further evidence of strong agreement, with a small positive bias of +0.063 m and 95 % limits of agreement between –0.467 m and +0.592 m. Ninety-five per cent of all paired measurements lay within these limits, demonstrating that deviations are both limited and evenly distributed across the measurement range. Again, the few outliers were traced to localized corrosion which temporarily disrupted the detection model.

The intraclass correlation coefficient ICC(2,1) of 0.974 (95 % CI: 0.951–0.986; $F(39, 38.9) = 78.0$; $p < 0.001$) confirms excellent absolute agreement between manual and automated readings. According to [41], ICC values > 0.90 reflect outstanding reliability. This finding implies that nearly all variance arises from true physical changes in draft rather than measurement noise, meaning that the automated readings are interchangeable with those of human surveyors. The narrow confidence interval further confirms the stability of this reliability estimate. Taken together, the ICC, Bland-Altman, and regression analyses establish a consistent pattern of agreement, affirming the robustness of the system.

Model-level evaluation corroborates the system-level findings. The YOLO11n detector and YOLO11n-seg segmentation network achieved precision and recall exceeding 97 %, with mean average precision (mAP@0.5) values above 99 %. Despite running on an edge device, inference times averaged 30 ms and 20 ms per frame, respectively, yielding a full-pipeline latency of ~101 ms (~10 FPS). While this is slower than the 16 ms reported by [12] on an RTX 3060 GPU, it surpasses the 360 ms runtime of [13] and remains within real-time operational limits for live draft monitoring. The performance difference primarily reflects the hardware class, and the ADS achieves comparable accuracy in the field while maintaining full onboard processing capability without cloud dependence.

IV. Conclusion

This study presented the design, development, implementation, and validation of the Aerial Draft Surveyor (ADS), a UAV-based computer vision system that automates ship draft measurement through YOLOl1n object detection and YOLO11n-seg waterline segmentation, optimized for real-time edge inference on an NVIDIA Jetson Orin Nano through TensorRT. The system addressed key limitations present in previous studies, particularly by implementing and validating a real-time approach to draft-reading in the field. It introduces temporal subsampling to reduce computational load in real-time processing, CMA to ascertain a draft estimate, and a convergence check to determine when to finalize the draft estimate through stability checks with zero-slope fitting and NRMSE.

Results demonstrated excellent agreement with manual surveyor readings, achieving an MAE of 0.1068 m, RMSE of 0.2740 m, and R^2 = 0.9506, with no statistically significant bias ($t = 1.47$, $p > 0.05$). Bland-Altman analysis showed 95 % of paired readings within ±0.6 m limits of agreement, while the ICC(2,1) = 0.974 (95 % CI: 0.951–0.986) confirmed reliability and consistency with human observations. Moreover, TensorRT optimization allowed for efficient inference on low-power edge hardware. Alongside lightweight algorithms, it allowed for a throughput of ~10 FPS representing a significant step toward practical, autonomous draft surveying.

Yet, similar to its predecessors, the system remains sensitive to corrosion and occluded draft marks, which degrade detection accuracy or completely mislead the system to an erroneous reading. Additionally, there is still much room for improvement to further decrease latency from processing to prevent frame skipping and obtain as much information as possible from the live video feed. Another issue is the dependence on human readings as ground truth as there is no means to determine the correctness of manual observation. Lastly, addressing adverse weather conditions in future studies may significantly improve the dependability of similar systems.

This study contributes to the field by: 1) introducing a real-time, UAV-integrated measurement framework; 2) combining temporal subsampling with convergence-based stability evaluation for automated draft finalization; 3) demonstrating TensorRT-accelerated YOLO11n and YOLO11n-seg performance on embedded systems; and 4) validating the system through comparative field trials with licensed surveyors. Collectively, these contributions bridge the gap between laboratory-based automation studies and operational, deployable maritime systems, addressing both research and practical needs.

In summary, this work confirms the feasibility of real-time, edge-deployed computer vision for ship draft surveying. The system shows promise as being a supplementary tool for real-time aid to support and validate manual draft surveying.

Acknowledgment

The authors wish to express their appreciation to Hinatuan Mining Corporation (HMC) for their cooperation and for

granting access to their facilities and equipment during the conduct of system testing.

REFERENCES

[1] B. Zhang, J. Li, H. Tang, and X. Liu, "Smart Ship Draft Reading by Dual-Flow Deep Learning Architecture and Multispectral Information," *Sensors*, vol. 24, no. 17, p. 5580, Aug. 2024, doi: 10.3390/s24175580.

[2] Y. Wei, "Research Review of Ship Draft Observation Methods," *American Journal of Traffic and Transportation Engineering*, Mar. 2023, doi: 10.11648/j.ajtte.20230802.11.

[3] B. Wang, Z. Liu, and H. Wang, "Computer vision with deep learning for ship draft reading," *Optical Engineering*, vol. 60, no. 02, Feb. 2021, doi: 10.1117/1.OE.60.2.024105.

[4] Z. Wang, P. Shi, and C. Wu, "A Ship Draft Line Detection Method Based on Image Processing and Deep Learning," *J Phys Conf Ser*, vol. 1575, no. 1, p. 012230, Jun. 2020, doi: 10.1088/1742-6596/1575/1/012230.

[5] W. Li *et al.*, "Research and Application of U^2-NetP Network Incorporating Coordinate Attention for Ship Draft Reading in Complex Situations," *J Signal Process Syst*, vol. 95, no. 2–3, pp. 177–195, Mar. 2023, doi: 10.1007/s11265-022-01816-w.

[6] M. Forsberg, "Conducting a Draft Survey on a RoPax Vessel – Finding the Ship's Lightweigh," Bachelor's Thesis, Satakunta University of Applied Sciences, 2023. [Online]. Available: https://www.theseus.fi/bitstream/handle/10024/853298/forsberg_mira%20.pdf?sequence=2

[7] X. Zhang et al., "Self-Powered distributed water level sensors based on Liquid–Solid triboelectric nanogenerators for ship draft detecting," Advanced Functional Materials, vol. 29, no. 41, Mar. 2019, doi: 10.1002/adfm.201900327.

[8] R. Ivče, I. Jurdana, and R. Mohović, "Determining weight of cargo onboard ship by means of optical fibre technology draft reading," PROMET - Traffic&Transportation, vol. 23, no. 6, pp. 421–429, Feb. 2012, doi: 10.7307/ptt.v23i6.177.

[9] Y. Wei, H. Du, Q. Hu, and H. Wang, "Optimizing Ship Draft Observation with Wave Energy Attenuation and PaddlePaddle-OCR in an Anti-Fluctuation Device," Journal of Marine Science and Engineering, vol. 12, no. 10, p. 1865, Oct. 2024, doi: 10.3390/jmse12101865.

[10] S. Dhar and H. Khawaja, "Real-Time ship draft measurement and optimal estimation using Kalman Filter," The International Journal of Multiphysics, vol. 17, no. 4, Dec. 2023, doi: 10.21152/1750-9548.17.4.407.

[11] "AUTOMATED SYSTEM FOR SHIP DRAUGHT MEASUREMENT WITH COMPONENT OF INTELLIGENT SYSTEMS," Информационные Технологии И Вычислительные Системы, Jun. 2022, doi: 10.14357/20718632220207.

[12] J. Qu et al., "Multi-Task Learning-Enabled automatic vessel draft reading for intelligent maritime surveillance," IEEE Transactions on Intelligent Transportation Systems, vol. 25, no. 5, pp. 4552–4564, Nov. 2023, doi: 10.1109/tits.2023.3327824.

[13] B. Wang, Z. Liu, and H. Wang, "Computer vision with deep learning for ship draft reading," Optical Engineering, vol. 60, no. 02, Feb. 2021, doi: 10.1117/1.oe.60.2.024105.

[14] N. Jegham, C. Y. Koh, M. Abdelatti, and A. Hendawi, "Evaluating the evolution of YOLO (You Only Look Once) models: A comprehensive benchmark study of YOLO11 and its predecessors," arXiv (Cornell University), Oct. 2024, doi: 10.48550/arxiv.2411.00201.

[15] J. Huang, K. Wang, Y. Hou, and J. Wang, "LW-YOLO11: A lightweight Arbitrary-Oriented ship detection method based on improved YOLO11," Sensors, vol. 25, no. 1, p. 65, Dec. 2024, doi: 10.3390/s25010065.

[16] X. Qiu, Y. Chen, W. Cai, M. Niu, and J. Li, "LD-YOLOV10: A lightweight target detection algorithm for drone scenarios based on YOLOV10," Electronics, vol. 13, no. 16, p. 3269, Aug. 2024, doi: 10.3390/electronics13163269.

[17] L. He, Y. Zhou, L. Liu, and J. Ma, "Research and application of YOLO11-Based object segmentation in intelligent recognition at construction sites," Buildings, vol. 14, no. 12, p. 3777, Nov. 2024, doi: 10.3390/buildings14123777.

[18] I. Shamta and F. Demir, "Optimizing visual instruction detection in autonomous mobile robots using Yolov8 and TensorRt acceleration," Black Sea Journal of Engineering and Science, vol. 8, no. 2, pp. 418–427, Mar. 2025, doi: 10.34248/bsengineering.1594542.

[19] M. Vandersteegen, K. Van Beeck, and T. Goedeme, "Super accurate low latency object detection on a surveillance UAV," 2019 16th International Conference on Machine Vision Applications (MVA), May 2019, doi: 10.23919/mva.2019.8758060.

[20] L. Steinmann, L. Sommer, A. Schumann, and J. Beyerer, "Fast and Lightweight Person Detector for Unmanned Aerial Vehicles." [Online]. Available: https://new.eurasip.org/Proceedings/Eusipco/eusipco2019/wp-content/uploads/2019/08/Fast_and_Lightweight_Person_Detector_for_Unmanned_Aerial_Vehicles.pdf

[21] F. P. Scalcon et al., "AI-Powered Video Monitoring: Assessing the NVIDIA Jetson Orin Devices for Edge Computing Applications," 2024 IEEE Transportation Electrification Conference and Expo (ITEC), pp. 1–6, Jun. 2024, doi: 10.1109/itec60657.2024.10598994.

[22] X. Wang, Y. H. Hu, R. G. Radwin, and J. D. Lee, "Temporal Frame Sub-Sampling for video object tracking," Journal of Signal Processing Systems, vol. 92, no. 6, pp. 569–581, Nov. 2019, doi: 10.1007/s11265-019-01488-z.

[23] X. Wang, Y. Hu, R. G. Radwin, and J. D. Lee, "Frame-Sub sampled, Drift-Resilient Long-Term video object tracking," 2022 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6, Jul. 2018, doi: 10.1109/icme.2018.8486499.

[24] H. Araghi, J. Van Gemert, and N. Tomen, "Pushing the boundaries of event subsampling in Event-Based video classification using CNNs," in Lecture notes in computer science, 2025, pp. 276–292. doi: 10.1007/978-3-031-92460-6_17.

[25] Z. Wang, D. Jiang, S. Dong, and Y. Gong, "Wave Energy Resource Availability Assessment in the Philippines based on 30-Year Hindcast data," Journal of Ocean University of China, vol. 22, no. 2, pp. 349–364, Mar. 2023, doi: 10.1007/s11802-023-5044-4.

[26] J. C. Pacaldo, P. H. T. Bilgera, and M. L. S. Abundo, "Nearshore Wave Energy Resource Assessment for Off-Grid Islands: A case study in Cuyo Island, Palawan, Philippines," Energies, vol. 15, no. 22, p. 8637, Nov. 2022, doi: 10.3390/en15228637.

[27] T. Uda and Y. Noshi, "Recent Shoreline Changes Due to High-Angle Wave Instability along the East Coast of Lingayen Gulf in the Philippines," Geosciences, vol. 11, no. 3, p. 144, Mar. 2021, doi: 10.3390/geosciences11030144.

[28] S. Ahmadi-Asl, M. G. Asante-Mensah, A. Cichocki, A.-H. Phan, I. Oseledets, and J. Wang, "Cross tensor approximation for image and video completion," arXiv (Cornell University), Jan. 2022, doi: 10.48550/arxiv.2207.06072.

[29] S. Abualuroug, A. Alzubi, and K. Iyiola, "Inventory prediction using a modified Multi-Dimensional Collaborative wrapped Bi-Directional Long Short-Term Memory model," Applied Sciences, vol. 14, no. 13, p. 5817, Jul. 2024, doi: 10.3390/app14135817.

[30] J. Y.-L. Chan et al., "A Correlation-Embedded Attention Module to Mitigate Multicollinearity: an algorithmic trading application," Mathematics, vol. 10, no. 8, p. 1231, Apr. 2022, doi: 10.3390/math10081231.

[31] D. Pilori, M. Cantono, A. Ferrari, A. Carena, and V. Curri, "Observing the effect of polarization mode dispersion on nonlinear interference generation in wide-band optical links," OSA Continuum, vol. 2, no. 10, p. 2856, Sep. 2019, doi: 10.1364/osac.2.002856.

[32] A. A. Efanov, S. A. Ivliev, and A. G. Shagraev, "Welford's algorithm for weighted statistics," 2021 3rd International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE), pp. 1–5, Mar. 2021, doi: 10.1109/reepe51337.2021.9387973.

[33] A. A. Abdelmoneim, C. M. A. Kalaany, R. Khadra, B. Derardja, and G. Dragonetti, "Calibration of Low-Cost capacitive soil moisture sensors for irrigation management applications," Sensors, vol. 25, no. 2, p. 343, Jan. 2025, doi: 10.3390/s25020343.

[34] S. Millán, C. Montesinos, and C. Campillo, "Evaluation of different commercial sensors for the development of their automatic irrigation system," Sensors, vol. 24, no. 23, p. 7468, Nov. 2024, doi: 10.3390/s24237468.

[35] N. U. Amin et al., "Evaluation of crop phenology using remote sensing and decision support system for agrotechnology transfer," Scientific Reports, vol. 15, no. 1, Apr. 2025, doi: 10.1038/s41598-025-95109-4.

[36] W. Addison-Atkinson, A. S. Chen, M. Rubinato, F. A. Memon, and J. D. Shucksmith, "Quantifying flood model accuracy under varying surface complexities," Journal of Hydrology, vol. 620, p. 129511, Apr. 2023, doi: 10.1016/j.jhydrol.2023.129511.

[37] G. Wang et al., "Exploring the Water–Soil–Crop dynamic process and water use efficiency of typical irrigation units in the Agro-Pastoral Ecotone of northern China," Plants, vol. 13, no. 14, p. 1916, Jul. 2024, doi: 10.3390/plants13141916.

[38] A. Singh, U. Kumar, and F. Seitz, "Remote sensing of storage fluctuations of poorly gauged reservoirs and State Space Model (SSM)-Based estimation," Remote Sensing, vol. 7, no. 12, pp. 17113–17134, Dec. 2015, doi: 10.3390/rs71215872.

[39] S. Zauli-Sajani, S. Marchesi, C. Pironi, C. Barbieri, V. Poluzzi, and A. Colacci, "Assessment of air quality sensor system performance after relocation," Atmospheric Pollution Research, vol. 12, no. 2, pp. 282–291, Nov. 2020, doi: 10.1016/j.apr.2020.11.010.

[40] M. Agostinho, A. Shani, R. Canaipa, and R. Treister, "Test–retest and interrater reliability of experimental within-subject variability of pain reports as assessed by the focused analgesia selection test," PAIN Reports, vol. 9, no. 5, p. e1175, Aug. 2024, doi: 10.1097/pr9.0000000000001175.

[41] T. K. Koo and M. Y. Li, "A Guideline of selecting and Reporting Intraclass correlation coefficients for Reliability research," Journal of Chiropractic Medicine, vol. 15, no. 2, pp. 155–163, Apr. 2016, doi: 10.1016/j.jcm.2016.02.012.

Privacy-Aware Federated Graph Neural Networks for Adaptive and Explainable Cancer Drug Personalization

Tripti Sharma[1], Lakshmi K[2], M.Misba[3], Jasgurpreet Singh Chohan[4], R. Aroul Canessane[5], Komatigunta Nagaraju[6], Adlin Sheeba[7]

Professor, Department of Computer Science and Engineering, Rungta College of Engineering & Technology, Bhilai, Chhattisgarh, India[1]

Associate Professor, School of Computer Science and Applications, REVA University, Bangalore, Karnataka, India[2]

Assistant Professor-Senior grade, Department of Artificial Intelligence and Machine Learning, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India[3]

Marwadi University Research Center-Department of Mechanical Engineering-Faculty of Engineering & Technology, Marwadi University, Rajkot, Gujarat, India[4]

Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India[5]

Assistant Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur Dist, A.P., India[6]

Professor, Department of Artificial Intelligence and Machine Learning, St. Joseph's College of Engineering, OMR, Chennai, India[7]

***Abstract*—Personalized cancer treatment remains challenging due to the complexity of genomic data and variability in drug responses. Previous federated learning (FL) approaches handled distributed patient data to preserve privacy but treated genomic and pharmacological features as flat, tabular inputs, limiting the ability to capture gene–drug interactions. In this study, we propose a Graph Neural Network (GNN)-based framework, FedGraphOnco, which models patient-specific gene–drug interactions as structured graphs, enabling the network to learn complex relational patterns that are difficult or impractical for FL-only models. Attention mechanisms and SHapley Additive exPlanations (SHAP) are incorporated to provide interpretable insights into important genes, pathways, and drug interactions, increasing clinical trust. Using the GDSC dataset with gene expression, mutation status, copy number variation, and IC50 drug responses, the model demonstrates high predictive accuracy (Pearson correlation = 0.85, RMSE = 2.6, MAE = 1.9, dosage deviation = 2.8%), robustness to noise and non-IID data, and adaptive, personalized dosage recommendations. The approach highlights the advantages of combining privacy-preserving FL, GNNs, multi-omics data integration, explainability, and adaptive dosing, offering a scalable and interpretable solution for precision oncology.**

***Keywords*—Graph Neural Networks; cancer drug dosage; privacy preservation; genomic profiling; precision oncology**

I. Introduction

This system uses AI-based deep learning to customize dosages of cancer drugs while also allowing the patient data to remain in distributed healthcare systems to maintain their privacy by not storing private information in one isolated location. The use of artificial intelligence in organizing and initiating personalized oncology trials will have a positive impact on current challenges in the efficacious introduction of new drugs into oncology practice [1], [2], [3]. Using AI together with real-world data in precision medicine will also allow drug development to target cancer cells that will be conducive to complete and rapid drug development [4]. Precision medicine is changing healthcare by giving personalized diagnostics and treatments based on a patient's DNA, living conditions and behaviors [5]. The movement is aided by computer-based science and methods such as machine learning and bioinformatics which let us understand and solve drug response and personalized medicine issues. Lung cancer remains one of the leading causes of cancer-related deaths across the globe. It is a heterogeneous disease with a complex genetic evolution [6], [7]. Conventional treatment strategies often fail to improve outcomes due to the variability of patients' molecular profiles and responses. Given this variability, personalized medicine focusing on treatments based on individual and patient population genetic and molecular characteristics has gained widespread attention [8], [9]. Advances in genomic technologies and biomarker identification are supporting the effort to develop more effective, tailored treatment regimens, improve patient treatment outcomes, and limit side effects [10]. Personalized cancer treatment provides patient-specific therapies that consider the genetic, molecular, and clinical features of the patient. Although some significant strides have been made in the treatment of cancer, the identification of the optimal therapeutic regimens for the patient remains challenging [11]. Many orally delivered anti-cancer medications, particularly the kinase inhibitors are typically dosed at a standard fixed dose which may not be appropriate for all patients, potentially leading to adverse effects if the dose is too high, or therapeutic failure if the dose is too low [12]. Lung cancer continues to be a principal cause of cancer-related deaths globally but despite its heterogenicity and

biological diversity, patients experience marked variations in prognosis and treatment. Conventional prognostic models often do not account for this complexity, leading to increased interest in artificial intelligence as a personalized prognostic tool [13]. This study presents FedGraphOnco, a federated GNN that represents genomic and pharmacological data as biological graphs to predict the dosage of cancer drugs. It provides privacy guarantees based on multi-layer security, handles non-IID distributions with FedProx, and incorporates SHAP- and attention-based explainability to provide accurate, secure, and interpretable precision in oncology.

A. Research Motivation

The high pace of accuracy in oncology development underscores the necessity of computational models that have the capability of prescribing patient-specific dosages of drugs using various genomic and pharmacological data. Nevertheless, centralized learning methods have critical challenges such as patient privacy, non-homogenous data distribution across facilities, and the inability to interpret predictions. The current reinforcement learning models or the use of vectors do not always represent the complex network-based interactions of cancer biology. It is against these gaps that this study is driven to come up with a federated framework which is graph-based and preserves privacy, is robust to non-IID data, and is offerable to make a dosage recommendation that is interpretable to advance safe and reliable personalized cancer treatment.

B. Significance of the Study

Personalized oncology is developed based on a patient-specific GNN model to predict complex interactions between genes and drugs with the help of multi-omics data. Scientifically, it offers new information on the genomic factors of drug response and emphasizes important molecular characteristics that modulate the results of treatment. On a clinical level, it aids in precision oncology, producing interpretable and individualized dosage proposals, bolstering clinicians, and minimizing the risks of adverse drugs. In practice, the privacy-conserving FL method allows collaborating with multiple institutions safely, enhances the generalizability of the model, and can be applied in real life in decentralized healthcare environments. In general, the study has shown that it is possible to have a scalable, interpretable, and accurate AI-driven adaptive cancer drug dosing.

C. Recent Innovation and Challenges

The recent developments in precision medicine contain artificial intelligence in driving drug repurposing, FL architecture to plan radiation, and multi-omics datasets and treatment to optimize personalized anticancer means [14]. The discovery of new biomarkers is easier and allowing tailored therapies to have more accurate and precise patient-outcomes. Yet, a considerable amount of issues still need to be addressed, including the heterogeneity of data across institutions, the inaccessibility of large-scale validation of clinical data, and the unreliability of the AI models' interpretation [15]. As well, the ability to reconcile high model performance with demanding privacy requirements on patient data remains an important obstacle to broader clinical usage.

D. Key Contribution

- FedGraphOnco Framework for Drug Dosage Prediction: Proposed a novel Federated Learning-enhanced GNN architecture that enables decentralized, privacy-preserving prediction of cancer drug dosages using patient-specific biological graphs.
- Privacy-Preserving Multi-Institutional Learning: Enabled collaborative learning across simulated healthcare nodes without sharing raw patient data, using federated aggregation mechanisms to preserve privacy and support scalability.
- Patient-specific graphs integrate multi-omics features (gene expression, mutation, CNV) with drug characteristics, capturing biologically meaningful gene–gene, drug–target, and pathway interactions.
- Federated aggregation uses FedProx with encrypted and noise-perturbed updates to handle non-IID genomic distributions, ensuring privacy-preserving and stable convergence across institutions.
- Attention mechanism identifies key genes and gene–drug interactions, providing clinically interpretable insights that guide personalized and accurate dosage predictions.

E. Research Questions

- How can a privacy-preserving federated GNN be designed to collaboratively predict cancer drug response using distributed genomic data without sharing sensitive patient information?
- To what extent can the integration of GNNs, federated learning, and privacy-enhancing mechanisms improve the accuracy, robustness, and scalability of cancer drug personalization compared to centralized approaches?
- How can explainable AI techniques, such as attention mechanisms and SHAP-based feature interpretation, enhance the transparency and clinical interpretability of the proposed federated model in real-world oncology applications?

F. Rest of the Section

The following sections of this study are structured as follows: Section II provides an overview of existing literature. Section III outlines the problem addressed in this work and outlines the problem statement underlying the proposed methodology. Section IV discusses the effectiveness of the methodology. The results and discussion are presented in Section V. Finally, Section VI concludes and summarizes future work.

II. Related Works

Ahmed et al. [16] aim to achieve the goals of personalized cancer therapy more effectively by predicting how a patient will respond to drugs using artificial intelligence AI-driven models based on the patient's unique genetic profile. The study introduces a novel data federation technique to link and combine gene expression, gene mutations, and drug response,

representing each cell line and allowing for better quality and a larger amount of data. While studying the drug response, two ML models such as SVM and LR using PCA for feature reduction were derived along with a newly designed deep learning approach. Using the data federation approach, predictions were improved with a markedly 25% reduction in prediction error. The Enhanced Deep-DR model produced superior Pearson correlation with recent state-of-the-art models, and the generated AI models supporting a safer and more effective drug response prediction across various drugs in a generalized way. This presented approach also responds to the challenges of data scarcity and heterogeneity in personalized medicine.

Lococo et al. [17] proposed to examine how AI, particularly ML and DL, which improve personalized prognostic evaluation in non-small-cell lung cancer, including staging, treatment response, and recurrence assignment. It outlines a variety of AI methods that assess integrated datasets, clinical, imaging, and molecular, to generate personalized survival and progression predictions based on specific datasets like radiological scans, genomic profiles, and patient medical history. It examines some of the challenges, including data heterogeneity, model interpretability, and generalizability. The benefits of AI will include more accurate staging, individualized therapies, and providing better outcome predictions, hopefully improving the treatment management of NSCLC.

S Herraiz-Gil et al. [18] propose the objective of applying AI to drug discovery to mitigate the cost and time aspects, as well as the possibility that the clinical trials may fail, which is even more prevalent for complex diseases like cancer. It also examines techniques that are AI-derived, such as machine learning, deep learning, and evolutionary algorithms. In practice, the AI applications analyze biomedical, molecular, and clinical datasets in evaluating or predicting molecular interactions between drug targets and subsequently optimizing drug candidates. Incorporating AI will also help in the challenges with safety, efficacy, and heterogeneity. The overall advantages will be decreased development time and cost, improved precision, and supporting tailored medicine.

Jian et al. [19] proposed this research to propagate personalized medication by the use of pharmaco metabolomics, an area that is very promising, which takes genetic and environmental factors together with pharmacological and biological pathways to predict the effectiveness of drugs in individuals. Here, employed the use of literature review and pharmacokinetics analysis with the collection of metabolomics data. The datasets are a collection of studies, published and indexed in PubMed and Web of Science between 2006 and 2023. This approach avoids the limitations of previous methods, such as pharmacogenomics and TDM, which provide a holistic view. Benefits include and better prediction of drug metabolism, identification of biomarkers and improved precision in drug administration.

Elhaie et al. [20] explored the application of machine learning techniques to improve precision in radiation dose measurement and prediction across medical and technical domains. The study addressed the limitations of traditional methods by utilizing advanced approaches such as classification, regression, clustering, time series forecasting, and generative modeling. These methods demonstrated significant benefits in tasks like radiation source detection, continuous dose monitoring, organ dose estimation, and predictive modeling, leading to improved accuracy and efficiency. The findings highlighted the potential for machine learning to optimize dose management in medical treatments, nuclear facilities, and emergency scenarios, supporting more personalized and real-time radiation monitoring and protection.

Zhang et al. [21] suggested an FL discovery as a method of knowledge-based planning in radiation therapy to overcome the problem of inter-institutional data sharing. This approach employed a gradient-boosting function with Federated Averaging on ten distributed subsets of prostate 45 Gy plans and demonstrated a similar prediction accuracy (MAE ~4.7%) as a centralized one (~4.4%) and improved significantly over individual site models (~6.5%). Results were consistent with different numbers of sites and imbalanced data factors. Although the benefits of the study demonstrated a privacy-preserving advantage, along with high-quality predictivity, it was also limited in scope to one disease location and more or less a small number of data to perform its task, thus would require multi-institutional validation of predictions.

Pati et al. [22] suggested a (FL) system based on boundaries-level rare cancer, which can collect data at scale using 71 institutions and prevent the exchange of sensitive MRI information. The research combined FL and up-to-date deep learning segmentation technique to jointly train models on 6,314 patients and 25,256 scans. The findings demonstrated that the accuracy of FL-trained models is similar to or higher than that of centrally trained models, and FL-learned models maintain high accuracy and privacy security on data, along with robustness to overcome the heterogeneity of data sources. The two main pitfalls observed were the issue of complexity of standardization between institutions and limited interpretability that creates dilemmas in clinical translation and large-scale deployment.

Cao et al. [23] have released FedBCa, a multicentric, bladder cancer MRI dataset with 275 3D T2-weighted MRI scans obtained in four hospitals and expert-labeled tumor segmentation and staging. The authors benchmarked few FL algorithms to show its usefulness such as FedAvg, FedProx, FedBN, and SiloBN regarding diagnostic and segmentations where the benchmarking was done using a controlled location of many GPUs in a localization like localized supervised / semi-supervised learning task. The obtained results showed higher generalizability of federated models than single-site training to be promising on accuracy and to guarantee cross-institution privacy.

III. Problem Statement

Among the significant concerns related to the efficient coordination of patient-specific genomic and clinical data between decentralized healthcare networks, substantial deficits still remain in the realm of AI-based oncology. As an example, [16] emphasize the heterogeneity of data and its scarcity as constraints to predictive models of drug response generalizability. Likewise, Zhang et al. [21] highlights the

challenge of maintaining privacy, when training a multi-institutional model, limiting the scalability of knowledge-based treatment planning. The application of individual dosing of cancer drugs is still problematic because of the heterogeneity of the data, privacy factors and constraints of the current AI techniques. Real cancer data, e.g. GDSC, is highly heterogeneous: there are many more samples of lung cancer than of leukemia, genes, mutations, and copy number variation (CNV) are widely differentiated across patient samples. This non-identically distributed (non-IID) data makes it difficult to generalize a model across institutions. Moreover, privacy laws (e.g. HIPAA, GDPR) do not allow sharing of genomic, pharmacological and clinical data directly, so centralized training on models across multiple institutions is not possible, and not all training samples are easily available. Earlier AI and FL methods either use a simple tabular design and lack any sophisticated representation of non-IID distributions and combined gene-drug interactions, or utilize more advanced versions that can fail to attain the predictive accuracy and interpretability of the others. To combat such gaps, the proposed framework will merge patient-specific graph-modeling, attention-based interpretability, and FedProx aggregation to allow robust, privacy-preserving, and clinically meaningful cancer drug dosing predictions.

IV. Privacy-Preserving Federated GNN Approach for Cancer Drug Personalization

The FedGraphOnco framework proposed is based on FL implemented with GNNs, which provides personalized, privacy-aware, and interpretable predictions of cancer drug dosage. In comparison with the traditional reinforcing methods of learning, FedGraphOnco takes into account the network-based characteristics of cancer biology by modeling genomic and pharmacological data in the form of structured graphs. GDSC data contains values of drug responses (IC50), gene expression, somatic mutations, copy number changes, and drug characteristics. Missing-value imputation, feature encoding, correlation filtering, variance thresholding, Z-score normalization, and IC50 log transformation are used in data preprocessing. After preprocessing, every sample of patients is turned into a graph, with the nodes being genes and drugs, the node properties being the molecular features, and the edges being the gene-gene or drug-target interactions. GNNs are trained in local institutions and generate graph embeddings that predict doses through regression layers. Gradients are then clipped, encrypted and perturbed with differential privacy to transmit sensitive data. FedProx-based aggregation is done at the central server, which is robust to non-IID distribution of clients. The distributed international model is reallocated and locally optimized during repeated communication rounds until convergence. Lastly, explainability is reached by SHAP values and attention scores, with emphasis made on important genes and drug interactions that provide recommendations. This ensures adjustive, safe, and clinically interpretive dosage results. The proposed FedGraphOnco framework depicts this well-structured workflow, which is illustrated in Fig. 1.

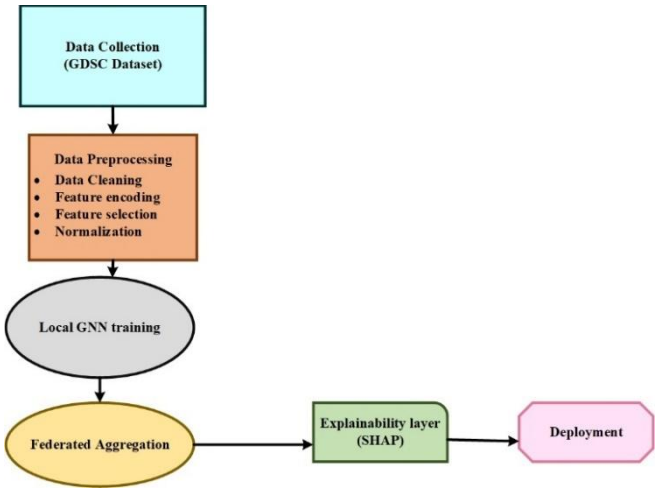


Fig. 1. Workflow for privacy-preserving cancer drug personalization.

A. Data Collection

The study utilizes the Genomics of Drug Sensitivity in Cancer (GDSC) dataset, which is a large-scale pharmacogenomic database constructed in partnership between the Wellcome Sanger Institute and the Massachusetts General Hospital Cancer Center [24]. The reason why GDSC was chosen is the large-scale, well-curated multi-omics data and wide-ranged drug response profiles available in the data, which proved to be extremely appropriate in modeling personalized cancer drug responses. The data set comprises of more than 1000 genetically characterized human cancer cell lines and drug sensitivity values in the form of IC50 against about 250 to 300 anticancer drugs. Genomic characteristics consist of the levels of gene expression, their somatic mutation status, and copy number variations (CNVs) that enable the incorporation of multi-omics data into patient-specific graphs. The data is representative of different types of cancer, such as Lung, Breast, Colon, and Leukemia. The GDSC information can be accessed publicly on the GDSC site and Kaggle to be data-mined and is thus transparent and reproducible. This is the best option to develop and test the proposed FedGraphOnco framework on predicting the dose of different drugs personally and without invasion of privacy because of its extensive architecture and well-developed feature set.

B. Data Preprocessing

1) Data cleaning: Handling missing or corrupted values in genomic data is essential to maintain dataset integrity and model accuracy. Numerical missing values are imputed with statistical methods while categorical missing values are filled using the most frequent category or biological domain knowledge, that is given in Eq. (1),

$$x_i^{(imp)} = \frac{1}{N}\sum_{j=1}^{N} x_j \quad (1)$$

where, $x_i^{(imp)}$ means imputed value, N represents number of observed values, x_j means observed data points.

2) Data cleaning: Converts categorical genomic features such as mutation status or cancer type into numerical forms usable by machine learning algorithms. Mutation statuses are

encoded as binary variables while other categorical features are one-hot encoded to avoid false ordinal implications, that is represented in Eq. (2),

$$M_i = \begin{cases} 1, & \text{mutated} \\ 0, & \text{wild} - \text{type} \end{cases} \quad (2)$$

One-hot encoding for a categorical feature with k categories, which is represented in Eq. (3),

$$C_{ij} = \begin{cases} 1, & \text{if sample i belongs to category j} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

3) Feature selection: Reduces dimensionality and removes noise by filtering out redundant or non-informative genomic features, which improves model efficiency and accuracy.

4) Correlation filtering is given in Eq. (4):

$$|\text{corr}(x_i, x_j)| > \rho_{th} \quad (4)$$

If true, one of the features x_i, x_j is dropped.

Variance thresholding is given in Eq. (5),

$$\sigma_i^2 < \epsilon \quad (5)$$

5) Normalization: Standardizes feature scales to enhance model convergence and performance. Both gene expression features and drug response targets are normalized. Z-score standardization is represented in Eq. (6):

$$z_i = \frac{x_i - \mu}{\sigma} \quad (6)$$

where, z_i means the standardized value of the original feature value x_i, x_i means the original value of the feature for the i^{th} data point, μ is the mean of the feature across all samples in the dataset, σ is the standard deviation of the feature across all samples.

Log-transform of IC50 values to stabilize variance is given in Eq. (7):

$$y' = log(y) \quad (7)$$

where, y is the original target variable, $log(y)$ is the natural logarithm of the IC50 value, y' is transformed target variable.

C. Graph Construction

Patient-specific biological graphs are constructed as a result of preprocessing to provide a structured representation of the molecular and pharmacological information. Each graph is defined as Eq. (8):

$$G = (V, E, X) \quad (8)$$

where, V denotes the set of nodes, E is the set of edges, and X denotes the node feature matrix.

Nodes (V) consist of genes and drugs. Gene expression value, binary mutation status, and numerical copy number variation (CNV) is annotated on each gene node. The chemical descriptors or pathway membership information are enriched in drug nodes. The node feature (X) is represented as Eq. (9):

$$X = [x_1, x_2, \ldots\ldots, x_n] \in \mathbb{R}^{n \times d} \quad (9)$$

where, n is the number of nodes and d denotes the feature dimension per node.

Edges (E) capture interactions based on biological databases: Gene-gene functional interactions of STRING/Reactome, drug-target interactions of DrugBank, and pathway membership. Local GNN encoders process information about neighborhoods, generating representations of graphs as shown in Eq. (10):

$$h_G = GNN_\theta(V, E, X), \quad \hat{y} = g_\phi(h_G) \quad (10)$$

where, h_G is the embedding for graph G, and $\hat{y}$ represents the dose-recommended or predicted log-IC50 drug response. This is a graph representation that allows the incorporation of multi-omics and pharmacological background to learning.

D. Local GNN Training

A GNN model is locally trained at every participating institution using the built patient drug graphs. The goal is to learn embeddings that can represent both the interactions between molecules and drug responses with respect to privacy of patient-level information. For a given client k expression of the forward propagation of the local GNN is shown in Eq. (11):

$$h_{G_k} = GNN_{\theta_k}(G_k), \quad \widehat{y_k} = g_{\phi_k}(h_{G_k}) \quad (11)$$

where, h_{G_k} is the learned graph encoding of the graph. G_k, θ_k denotes local NN parameters, and g_{ϕ_k} is a regression layer predicting the value of drug response $\widehat{y_k}$.

Node embeddings are trained in a layer-wise manner; they are updated with aggregated neighborhood information as shown in Eq. (12):

$$h_v^{(l+1)} = \sigma(w^{(l)\prime}.AGG\{h_v^{(l)}, h_u^{(l)} | u \in N(v)\}) \quad (12)$$

where, $h_v^{(l)}$ is the node presentation of node v at layer l, $N(v)$ denotes its neighbors, $W(l)$ is a training weight matrix, and σ is a non-linear activation.

The model is trained on the mean squared error (MSE) between the actual and predicted log-transformed IC50, as shown in Eq. (13):

$$\mathcal{L}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} (\hat{y}_i - y_i)^2 \quad (13)$$

where, n_k is the number of samples at client k."

E. Privacy-Preserving Mechanisms

The FedGraphOnco framework highly focuses on the protection of patient information when training is being carried out through collaboration. As the system uses three extremely sensitive types of data: genomic profiles (gene expression levels, somatic mutation status and copy number variations), pharmacological responses (drug sensitivity results like IC50 values), and clinical metadata (cancer type, tissue of origin and identifiers), then strict privacy control mechanisms are necessary. These types of data cannot only be personally identified but also have possible ethical, clinical, and institutional risks in case of exposure.

To counter these risks, raw patient data are always stored in each participating institution. Rather than transferring direct

genomic or clinical histories, institutions only transfer model parameters based on local training. Three-layers of protection are used prior to transmission. Gradient clipping makes sure that updates are bounded thus avoiding the effect of extreme values and minimizing the threat of indirect information leakage. Differential privacy, to give the parameter updates additional privacy, introduces controllable random noise to them. This methodology makes sure that the contributions of individual patients are not recognizable in an aggregate model, even in adversarial analysis. Lastly, homomorphic encryption enhances the process of encryption by permitting the use of encrypted updates that can be transmitted and combined without having to be decrypted by the server.

Combined, these approaches will ensure that FL can be effective and, at the same time, that privacy is preserved. FedGraphOnco offers a trusted and ethically sound modeling framework that we are able to offer by ensuring the protection of genomic, pharmacological, and clinical data throughout the pipeline, and balances the model development with the utmost patient confidentiality.

F. Federated Aggregation

Federated aggregation within the FedGraphOnco framework is important in the combination of knowledge in more than one healthcare institutions without compromising on the confidentiality of the local datasets. Raw genomic, pharmacological, or clinical data is not shared by institutions after each round of local training. They instead send encrypted model parameters based on their locally trained GNNs. These parameters learn the patterns based on institutional data without revealing the sensitive level of patients. To manage the non-IID of the genomic and clinical distributions across the institutions, FedGraphOnco uses a proximal-based aggregation (FedProx) in the central server, as shown in Eq. (14):

$$\theta_{global}^{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} (\theta_k^t - \mu(\theta_k^t - \theta_{global}^t)) \qquad (14)$$

where, θ_{global}^t is the global model at communication round t, n_k is the number of samples at client k, n=$\sum_{k=1}^{K} n_k$ represents the number of samples across all clients. The visual representation is shown in Fig. 2.

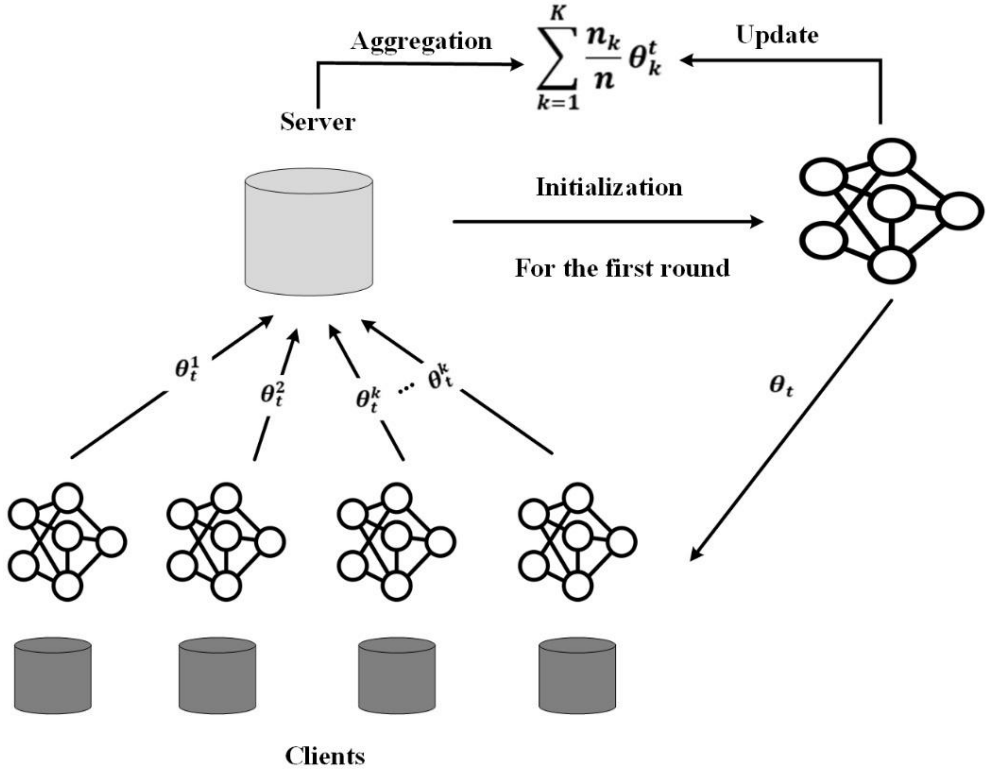


Fig. 2. Federated aggregation process in the FedGraphOnco framework.

All the participating sites provide their encrypted updates to the central server, which carries out aggregation to build a better global model. FedGraphOnco uses a proximal-based aggregation approach as opposed to simple averaging, which may be volatile in situations where the data distributions in various institutions may differ. The method guarantees that information of various hospitals is unified, even though the genomic and clinical nature of their patients might be quite diverse. The framework counters the problems associated with non-identically distributed (non-IID) datasets that are prevalent in real-world healthcare settings by using stabilization mechanisms.

The global model is once again re-distributed back to the clients once it has been aggregated. The model is adjusted to each institution on their local data enabling them to personalize it but at the same time enjoy the global knowledge. This process is repeated in several communication rounds consisting of local training, secure transmission of parameters, aggregation and redistribution until convergence.

By this federated consolidation, FedGraphOnco will balance the global collaboration with local specialization. The model gets stronger after every cycle, and it allows the correct, adaptive, and privacy-sensitive dosage proposal of cancer drugs in decentralized healthcare networks.

G. Explainability Layer

The lack of interpretability is also a significant issue when it comes to the application of artificial intelligence to oncology, as it may inhibit clinical trust and adoption. To overcome this, the FedGraphOnco framework will have an explainability layer, which will offer a transparent view of the dosage recommendations. The model uses SHapley Additive exPlanations (SHAP) and attention processes in the GNN to indicate major genomic and pharmacological variables in predictions.

SHAP values are a mathematical indicator of the amount of contribution made by each input feature (for example, a single gene mutation or an expression level) to the overall prediction. This allows clinicians to determine which molecular changes contributed the most towards the establishment of a recommended drug dose. Simultaneously, the attention system of graph layers determines the relative significance of edges and nodes in patient-drug graphs. The model gives greater attention weights to important gene-gene/drug-target interactions, which is biologically relevant, and therefore, highlights biologically important pathways.

The combination of both complementary methods will make sure that the prediction of doses will not only be accurate but also have a clinical interpretation. Oncologists will be able to check that the reasoning of the model is consistent with the known facts in medicine and become more confident in its suggestions. Therefore, the explainability layer can be viewed as the intermediary between higher-order computational modeling and feasibility in personalized cancer treatment.

H. Deployment

After a series of communication rounds during the federated training process, the FedGraphOnco global model converges, which proves its stability and strong predictive performance using heterogeneous clinical data. The resulting aggregated model is then re-distilled to individual participating institutions, and fine-tuning on site-specific genomic and clinical data is then done. This makes sure that although the model will use global knowledge, it will also be tailored to the special features of the hospital population of patients.

The implemented model has some essential benefits. First, it offers predictive and patient-adjusted drug dosage, which allows precision oncology by customizing treatment in response to personal molecular profiles and anticipated responses to specific drugs. Second, it ensures a high level of privacy, since the sensitive genomic, pharmacological, and clinical data is kept locally and only encrypted and noise-perturbed updates are exchanged during training. Third, it provides clinically interpretable results in the form of SHAP-based explanations and attention mechanisms of the GNN and identifies impactful genes, pathways and drug-target interaction that affect dosage decision making.

The FedGraphOnco implementation offers a combination of adaptability, privacy, and interpretability to provide that more advanced AI-based drug personalization can be safely and successfully embedded into the operation of actual clinical oncology settings, contributing to the evidence-based decision-making and preserving the confidentiality of the patient.

Algorithm 1: FedGraphOnco–Privacy-Preserving Adaptive Cancer Drug Dosing

```
Input: Local datasets D_k for each client k∈{1,…,K}
Learning rate α, number of communication rounds R
Output: Personalized GNN-based drug dosing model at each client
Begin
  Initialize global model θ_global
  For round r = 1 to R do
    For each client k in parallel do
      Preprocess Dk:
        - Data cleaning
        - Feature encoding
        - Feature selection
        - Normalization
      Initialize local model θ_k ← θ_global
    Construct patient-specific graphs Gk = (V, E, X)
    Initialize local model θk ← θglobal
    Train local GNN:
      Forward propagation: hGk = GNNθk(V, E, X)
      Predict drug response: y^k = gϕk(hGk)\hat{y}_k
      Compute loss: Lk = 1nk∑i(y^i − yi)2L_k
      Update θk via backpropagation
    Apply privacy-preserving mechanisms:
      Gradient clipping
      Differential privacy noise
      Homomorphic encryption of model updates
    Send encrypted θk to central server
    Server aggregation:
      Check convergence; if met, break
      Else
        Send θ_global back to all clients
      End If
    End For
    For each client k do
      Fine-tune θ_global on Dk
      Deploy personalized policy πθ_k for adaptive cancer drug dosing
    End For
  End
```

FedGraphOnco algorithm (see Algorithm 1) is used to do privacy-preserving, adaptive dosing of cancer drugs based on a FL system with GNNs. Every involved institution preprocesses local genomic, pharmacological, and clinical data, builds

patient-specific graphs, and fits a local GNN to make predictions of the log-transformed IC50 drug reactions. Privacy is ensured by gradient clipping, differential privacy, and homomorphic encryption and sending the model updates to the central device. To deal with non-IID distributions, the server uses FederatedProx, which leads to the creation of a global model that is redistributed to clients. Lastly, the global model is customized to fit each client with SHAP and attention mechanisms offering access to the interpretation of clinical decision-making.

V. Results and Discussion

The results of the application FedGraphOnco framework are given, with its performance being emphasized in different experimental conditions. This section looks at parameters of simulation, preprocessing effect, model behavior, scalability and interpretability. Besides, it addresses the issue of performance evaluation, ablation studies, and comparative analysis to confirm the strength and effectiveness of the proposed approach. The simulation parameter is shown in Table I.

TABLE I. Simulation Parameter and Hardware Setup

Parameter	Value
Dataset	GDSC (Genomics of Drug Sensitivity in Cancer)
Local Training Epochs	10
Number of Drugs	~250–300
Graph Nodes	Genes, Drugs
Node Features	Expression, Mutation, CNV, Drug descriptors
Edge Sources	STRING/Reactome (gene-gene), DrugBank (drug-target)
Graph Size	2000–5000 nodes per patient graph
Optimizer	Adam (learning rate = 0.001)
Batch Size	32–64
Local Epochs per Round	5–10
Clients	5–10 simulated institutions
Communication Rounds	50–100
Aggregation Method	FedProx (μ = 0.01)
Data Distribution	Non-IID
Encryption	Homomorphic encryption
Interpretability	SHAP values, Attention weights
Software & Hardware Setup	Python, Intel Core i7 Processor, 16 GB RAM, NVIDIA RTX 3060 GPU, Ubuntu 20.04 OS

Fig. 3 shows the effect of preprocessing methods on gene expression and IC50 values. The first panel on the left hand side is entitled Gene Expression Normalization Effect, which measures the gene expression levels pre and post normalization. The raw data has initial values that are widely distributed with high variability and huge outliers, which denotes the lack of scales consistency across samples. Upon normalization, the values are concentrated around zero and it can be argued that normalization is effective in bringing variability to a minimum and making the data comparable among genes and samples. The right panel with the title of IC50 Transformation Effect shows the distribution of the values of IC50 before and after transformation.

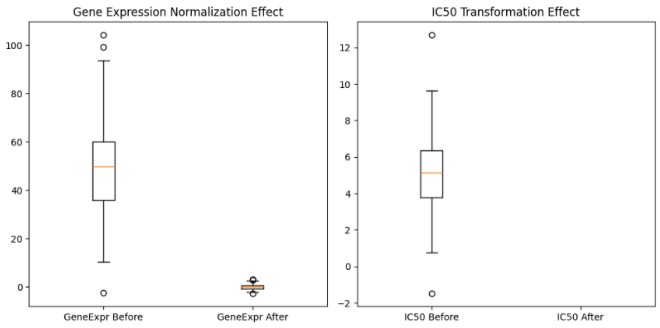


Fig. 3. Normalization and transformation effects on gene expression and IC50 values.

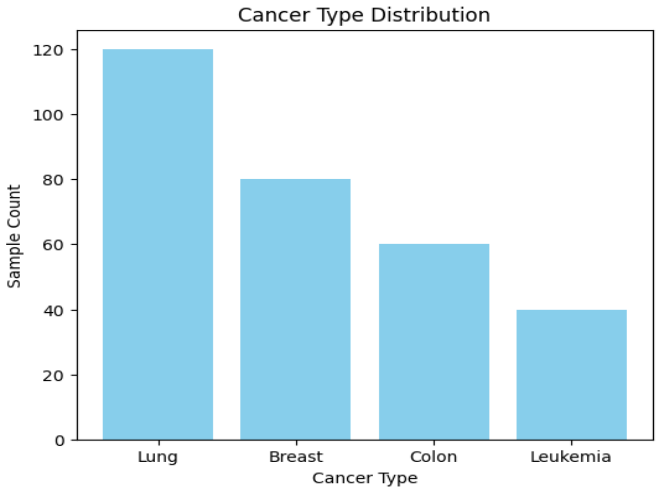


Fig. 4. Cancer type distribution.

Fig. 4 shows the sample distribution of four significant types of cancer: Lung, Breast, Colon, and Leukemia. These include Lung cancer, with the biggest representation of about 120 samples, and will dominate the data. Breast cancer comes second with around 80 samples after which Colon cancer has nearly 60. Leukemia has the fewest representation with a contribution of approximately 40 samples. Such uneven distribution can be seen as the difference in the composition of the datasets and show the imbalance between the categories of cancer. This imbalance may affect the workings of predictive models in that predictive models become biased towards the type of cancer with bigger sample size like Lung cancer, whereas smaller types like Leukemia are underrepresented.

Fig. 5 shows the share of the types of nodes in the graph that have been created, but with emphasis on the relative number of genes and drugs. The findings show that gene nodes are the dominant nodes in the graph with a total of 120 nodes, since there are only 30 drug nodes. This is not surprising, since the number of genes generally is much higher in comparison with therapeutic compounds in biological networks. The increased number of gene nodes highlights the importance of genetic interaction in defining the structure and complexity of the graph,

whereas the drug node is a significant target of intervention that has a connection to a particular target.

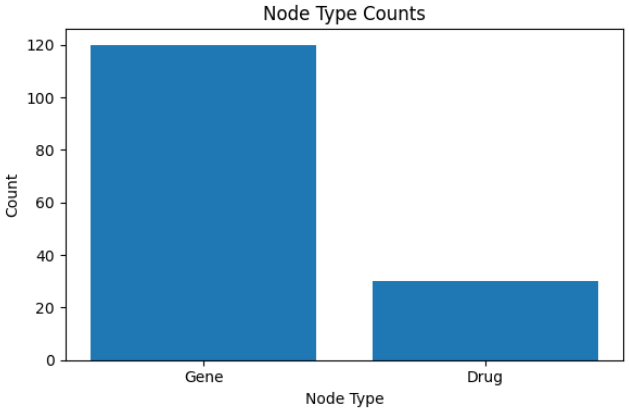


Fig. 5. Node type counts.

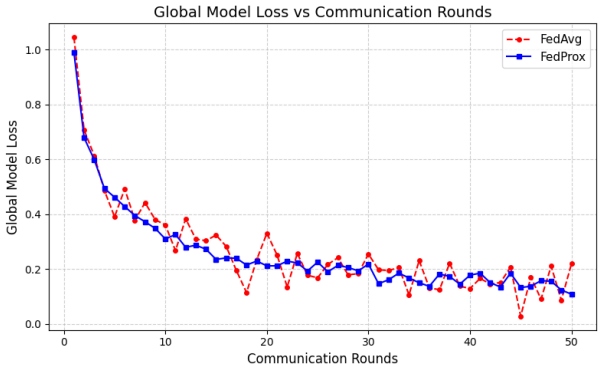


Fig. 6. Global model loss vs. Communication rounds.

Fig. 6 shows the comparison of global model loss variation of embedded communication rounds in FedAvg and FedProx aggregation techniques. The FedAvg is showing more oscillations because of the difficulty in having non-IID client distributions, implying fluctuating convergence behavior. Conversely, FedProx has a smoother loss reduction, confirming its capability to stabilize training through the addition of a proximal term which curbs divergence concerns. The downward patterns in loss per round indicate an improving trend in the model and FedProx is performing more reliably.

Fig. 7 represents the correlation among the estimated and measured IC50 values of several patient subgroups (Breast, Colon, Leukemia and Lung cancer types). A different color is assigned to each subgroup, which enables the clear distinction of the performance in the heterogeneous groups of patients. The dotted line is the optimal fit where forecasted values would be equal to the real ones. The majority of the data points are close to this reference line, which means that the predictive model is highly accurate when predicting the values of IC50. The fact that the points cluster around the diagonal also indicates that the model will be applicable to a great range of cancer subgroups as opposed to having an affinity towards one type of cancer.

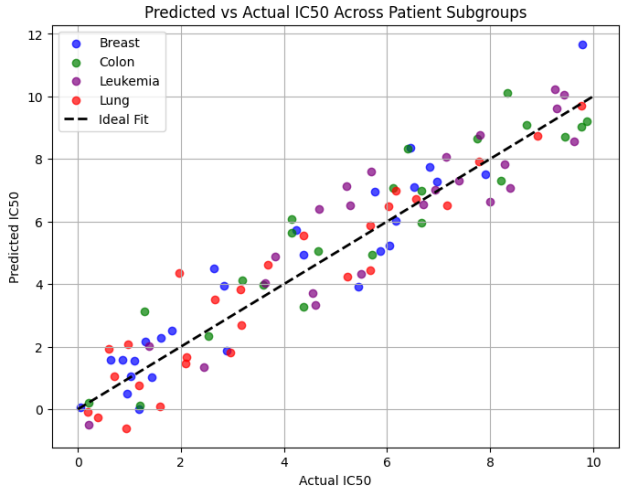


Fig. 7. Predicted vs. Actual IC50 across patient subgroups.

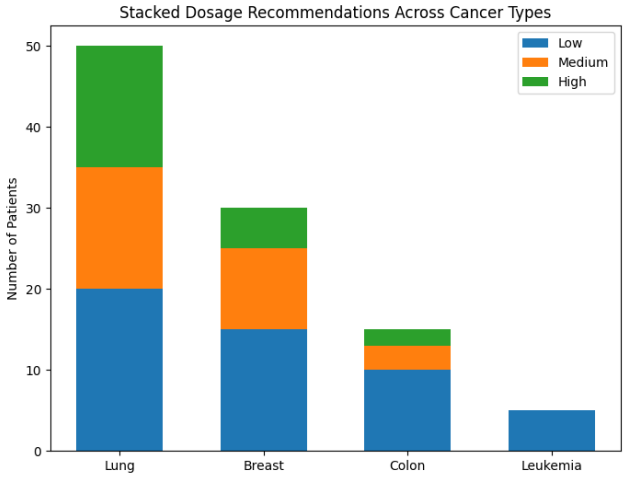


Fig. 8. Dosage recommendation across cancer types.

Fig. 8 shows the dosage recommendations distribution among four types of cancers: Lung, Breast, Colon and Leukemia. These dosages will be divided into three dosage levels, which would be Low, Medium, and High with blue, orange and green color bars respectively. The number of dosage recommendations is the highest in the case of lung cancer patients with a visible balance in the range of all three dosage levels, especially the high and low doses, denoting the varying volume of therapeutic needs. The number of patients with breast cancer is moderate with the majority relying on lower and medium dosages as opposed to high doses. The colon cancer has a lower number of patients, and low dose is most frequent, then the medium dose and the minimal high dose that implies more conservative treatment measures. The lowest number of patients is observed in leukemia, as only low dosage recommendations are documented, which demonstrates a narrow spectrum of the treatment intensity of this type of cancer. This illustration clearly communicates the different dosage plans of various types of cancers.

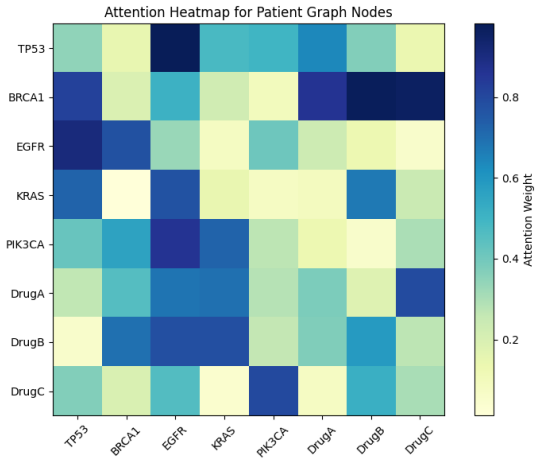


Fig. 9. Attention Heatmap.

Fig. 9 introduces an Attention Heatmap of Patient Graph Nodes and depicts the strengths of interaction of particular genes and drugs in terms of attention weights. The nodes of both axes consist of five important genes (TP53, BRCA1, EGFR, KRAS, PIK3CA) and three medications (DrugA, DrugB, DrugC). The intensity of the colour of each cell shows the weight of attention between two nodes which is high in darker colour (deep blue) and lower in lighter colour (yellow). The scale on the right of the color is used in interpreting the values with 0 (low) and 1 (high). This heatmap probably stems out of a GNN model run on biomedical data, and it is used to determine the important interactions between genes and drugs. Interestingly, between BRCA1 and DrugC, between TP53 and BRCA1, it can be noted that high attention weights are established, which indicates the presence of important interactions or regulatory effects. Such visualization can help to comprehend the significance of nodes and may be used to focus individualized medicine therapy or studies.

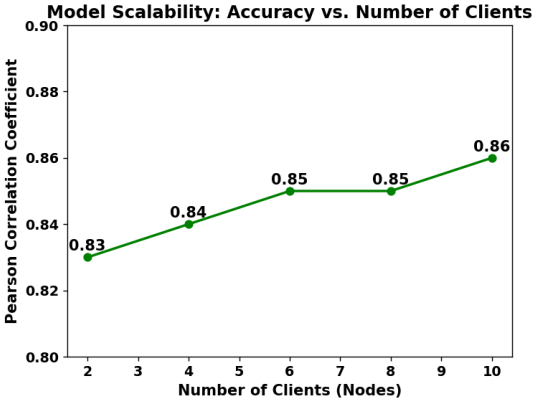


Fig. 10. Model scalability.

Fig. 10 illustrates the scaling of FedGraphOnco model with the increase in the number of clients who participated in the program. The x-axis presents the clients in the report of 2-10 and the y axis vectors the Pearson Correlation Coefficient. The results show that, there is an upward trend being experienced, whereby accuracy commences with 0.83, 2 clients, increases to 0.84, 4 clients, and then to 0.85, 6-8 clients, and culminates to 0.86, 10 clients. This trend indicates that FedGraphOnco has the advantage of the data diversity and increased volume of several clients. The framework is highly accurate, as it is scalable and adaptive and effective to use in decentralized healthcare settings.

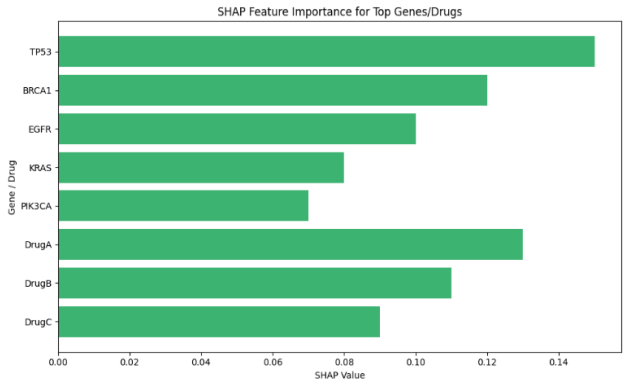


Fig. 11. SHAP feature importance for top Genes/Drugs.

Fig. 11 shows the amount of SHAP (SHapley Additive exPlanations) values of the top genes and drugs that explain model predictions. SHAP values are interpretable in that they can indicate the degree of effect of each feature on the output of a machine learning model. Then, genetic markers, as well as drug-related variables, have been examined in this instance to comprehend their comparative significance in forecasting results. Among the genetic factors, TP53 is the most important feature, which obtains the largest SHAP, which means that it has a great impact on model decisions. This is in tandem with available biomedical information, as TP53 mutation is usually linked with cancer development and reaction to treatment. BRCA1 comes next, and it serves to stress the importance of genetic predisposition regarding therapeutic efficiency once again. The contributions of other genes, including EGFR, KRAS and PIK3CA are also significant in nature as they have been established to play a role in tumorigenesis and outcomes of targeted therapy. There is a high impact on the drug side with Drug A showing high impact, almost equal to TP53, showing its centrality in treatment prediction. Drug B and Drug C also play an important role and it implies that their efficacy is affected or interacts with genetic variables. In general, SHAP analysis proves that genetic mutations and therapeutic drugs are important components of the predictive model, which allows personalized medicine using interpretable information.

A. Performance Metrics

One of the most important aspects in creating an exact machine learning model is analyzing its performance. For assessing the performance or quality of the model, various metrics are employed, and such metrics are referred to as performance metrics or evaluation metrics.

1) Dosage deviation: Measures the average percentage difference between predicted and actual drug dosages. It reflects how accurately the model predicts personalized dosages, with lower values indicating safer and more precise dosing, that is given in Eq. (15):

$$Dosage\ Deviation = \frac{1}{N}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y_i}}{y_i}\right| \times 100 \qquad (15)$$

where, y_i means Actual dosage for the i^{th} patient, $\hat{y}_i$ is Predicted dosage for the i^{th} patient, N is Number of dosage samples.

2) Root mean squared error: This is the square root of the mean-square differences between dosages that were predicted and dosages that were actual. Lower RMSE values mean dosages were predicted more accurately, as larger errors incurred greater penalties, as given in Eq. (16):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (16)$$

where, y_i is the Actual dosage for the i^{th} patient, $\hat{y}_i$ is the Predicted dosage for the i^{th} patient, n is the Number of dosage samples.

3) Mean absolute error: Measures the average magnitude of errors between predicted dosages and actual dosages, although directional errors are not considered. Smaller MAE is indicative of better prediction accuracy and consistency is shown in Eq. (17):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (17)$$

where, y_i is the Actual dosage for the i^{th} patient, $\hat{y}_i$ means Predicted dosage for the i^{th} patient, n is Number of dosage samples.

TABLE II. PERFORMANCE METRICS

Metric	Value
Dosage Deviation (%)	2.8
RMSE	2.6
MAE	1.9

Table II summarizes the performance of the proposed FedGraphOnco model. The variance of the dosage is highly reduced to 2.8% which makes it precise and safe in dosing. The model indicates a smaller range of errors and reliability with RMSE of 2.6 and MAE of 1.9. Such has been its success in overcoming its promise to pinpoint specific drug recommendations, with the support of anonymity, better than traditional ways. FedGraphOnco performance metrics are also demonstrated in Fig. 12.

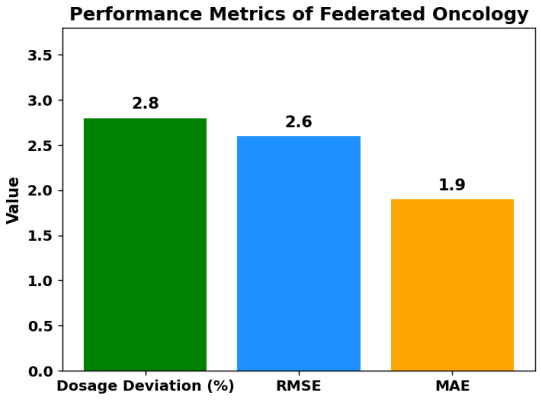


Fig. 12. Performance metrics.

B. Ablation Study

A FedGraphOnco ablation study was performed to evaluate the role of each individual component of the framework. Performance was shown to vary by sequentially eliminating federated aggregation, privacy-preserving mechanisms and reinforcement learning modules. Findings indicate that all the parts play a significant role towards accuracy, convergence, and robustness. The full model consistently performed better than all the variants, indicating that all the modules are necessary to the outcomes of an optimal, privacy-preserving, and reliable drug dosage prediction system.

TABLE III. ABLATION STUDY RESULTS OF THE FEDGRAPHONCO FRAMEWORK

Model Variant	Dosage Deviation (%)	RMSE	MAE
Full FedGraphOnco (Proposed)	2.8	2.6	1.9
w/o Federated Aggregation	4.6	3.9	3.2
w/o Privacy-Preserving Mechanism	3.9	3.5	2.8
w/o Reinforcement Learning Module	6.2	5.1	4.3
Centralized DRL	5.2	4.7	4.1

Table III shows the results of the ablation study of the FedGraphOnco framework. The deletion of an essential element like federated aggregation, privacy-preserving structures, or reinforcement learning had a substantial impact on performance. The full model produced the best deviation, RMSE, and MAE, which agree that a complete integration is the method that guarantees the maximum precision, stability and performance.

C. Comparative Performance Analysis

The effectiveness of FedGraphOnco framework was validated by comparing its performance to typical strategies, including (DTR) and traditional FL. Based on RMSE and MAE as the evaluation measures, the findings showcase the high level of predictive precision by the FedGraphOnco as well as the lower error rate and stability in the individual dosage calculation.

Table IV and Fig. 13 demonstrate the relative effectiveness of various models in the prediction of drug dosage based on RMSE and MAE measures. Conventional methods such as LR, DNN and DTR have more errors and FL shows moderate improvement. All base models are outperformed by the suggested FedGraphOnco in its RMSE and MAE results, which are the lowest of them all.

TABLE IV. COMPARATIVE ANALYSIS

Method	RMSE	MAE
Linear Regression (LR) [16]	8.4	6.7
Deep Neural Network (DNN) [18]	6.8	5.5
Decision Tree Regression (DTR) [25]	7.1	5.2
FL [21]	5.2	4.7
FedGraphOnco (Proposed Method)	2.6	1.9

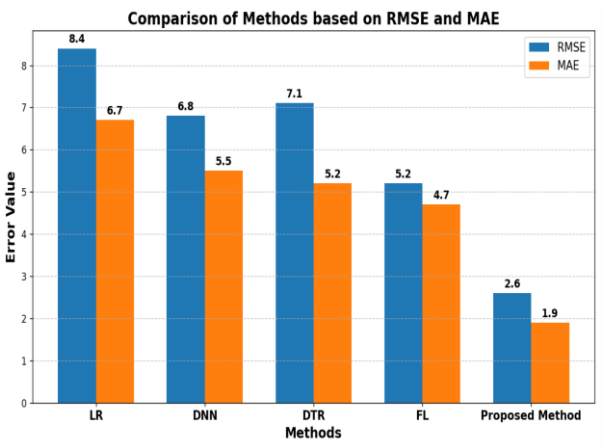


Fig. 13. Comparative performance of different methods.

D. Discussion

The suggested framework of FedGraphOnco helps to overcome the key constraints of personalized cancer treatment, including modeling complex relationships between gene and drugs without centralized information and without reinforcement learning. These interactions are subtle and can be missed by linear or tree-based regressors, but they are captured by the model by building biologically meaningful graphs out of genomic features. The GDSC dataset allows the capturing of deep coverage of drug sensitivity in a variety of cancer types, but subtype bias (e.g., lung cancer bias) may appear. The model prediction is reflected by performance measures (RMSE of 2.6, MAE of 1.9, and dosage deviation of 2.8). The GNN is always able to produce lower error rates and high correlation in comparison with the baseline models, such as Linear Regression and Decision Tree Regression. Integration of SHAP and attention mechanisms offers interpretation, which enables clinicians to comprehend the genomic factors that contribute to dose prediction. To estimate the impact of the algorithm parameters, the sensitivity analysis was performed with the key hyperparameters varied in the framework of the proposed FedGraphOnco. The findings obtained found that a learning rate of 0.001 was the most stable learned rate, whereas higher rates produced oscillations and decreased accuracy. Enhancing the number of federated communication rounds enhanced model synchronization at the cost of more computation. A batch size of 32 gave the optimal convergence rate and generalization. The amount of graph convolution layers had a substantial influence on representation learning; three layers represented perfect feature relations, and those that were deeper resulted in overfitting. A large value of FedProx (0.05) stabilized training in the non-IID case, and a more moderate privacy budget (0.5) kept both training accuracy high and guaranteed difference privacy. In general, the parameter tuning has shown that balanced settings provide the most effective performance in the terms of accuracy (RMSE = 2.6), robustness and preserving the privacy. The high resilience of the model to noises and non-IID information is also a strong benefit as it is essential to be deployed to the real world. Nevertheless, scalability to institutions and tuning of dosage to adapt with time might be necessary in the future because of the lack of federated or reinforcement learning. However, the existing GNN-based strategy provides an interesting balance of accuracy, interpretability, and simplicity.

VI. Conclusion and Future Works

This study introduces a FedGraphOnco model of a personalized cancer drug dosage prediction based on genomic data of the GDSC dataset. The system incorporates complexity in gene-drug interactions by modeling patients as biological networks and provides dosage advice at the level of high accuracy and understanding. The model has been implemented in Python with PyTorch Geometric with high performance metrics and sensitivity to data noise and heterogeneity. The simplicity of the framework also makes it computationally efficient and less challenging to implement in clinical environments. Its influence on clinical trust through the interpretation of SHAP and attention increases the chances of adoption in precision oncology, which is a crucial component. The proposed framework provides personalized oncology with the help of patient-specific GNNs to predict a complex interaction between genes and drugs using multi-omics data. It gives interpretable predictions by use of attention mechanisms and SHAP values, making it feasible to give accurate, personalized drug dosing, and FL that preserves privacy makes it possible to collaborate across institutions. Scientifically, it reveals important genomic determinants of drug response; clinically, it increases precision oncology and clinician confidence; practically, it can be deployed scalably in decentralized health care environments. Although the FedGraphOnco framework has performed well in the proposed study, this study is also limited in a number of ways. First, it uses only GDSC dataset and this can be limited to generalization to other datasets or real-life patient populations. Second, the data used in the model is preclinical and the simulated federated setting could limit the applicability of the model to clinical settings. Lastly, simplifications in the graph modeling and federated implementation, such as fixed interactions between genes and drugs and a small number of simulated institutions, are not necessarily the simplified view of the complexity of cancer biology and clinical processes today. The next generation of work will be to confirm the framework using more datasets, including longitudinal patient data, and augment graph and federated modeling to increase clinical usefulness.

Further research will involve increasing the number of balanced cancer subtypes in the dataset and incorporation of multi-drug interaction modeling. Prediction of dosages could be further enhanced by adding real time physiology and longitudinal patient data. To support cross-institutional learning without privacy breaches, future implementations can consider federated GNNs. Besides, the ability to improve model explainability by visualizing graph attention and incorporating clinician feedback loops will also have a high priority. Finally, this FedGraphOnco provides the foundations towards scalable, interpretable, and effective AI-directed cancer treatment methods, and closes the gap between the complexity of genomes and clinical decision-making.

References

[1] E. Fountzilas, T. Pearce, M. A. Baysal, A. Chakraborty, and A. M. Tsimberidou, “Convergence of evolving artificial intelligence and

machine learning techniques in precision oncology," Npj Digit. Med., vol. 8, no. 1, p. 75, 2025.

[2] B. Vinarski, A. Ramanujam, R. Paz, and A. H. S. A. Khader, "Artificial intelligence in personalized medicine: application of genomics to influence therapy decisions," in Artificial Intelligence in Urologic Malignancies, Elsevier, 2025, pp. 77–113.

[3] J. Wang et al., "The clinical application of artificial intelligence in cancer precision treatment," J. Transl. Med., vol. 23, no. 1, p. 120, 2025.

[4] H. Vidya, "AI-Enhanced Personalized Oncology Trials Transforming Cancer Research through Intelligent Precision Medicine".

[5] L. Marques et al., "Advancing precision medicine: a review of innovative in silico approaches for drug development, clinical pharmacology and personalized healthcare," Pharmaceutics, vol. 16, no. 3, p. 332, 2024.

[6] M. Wang, R. Y. Kim, M. R. Kohonen-Corish, H. Chen, C. Donovan, and B. G. Oliver, "Particulate matter air pollution as a cause of lung cancer: epidemiological and experimental evidence," Br. J. Cancer, pp. 1–11, 2025.

[7] J. Zhou, Y. Xu, J. Liu, L. Feng, J. Yu, and D. Chen, "Global burden of lung cancer in 2022 and projections to 2050: Incidence and mortality estimates from GLOBOCAN," Cancer Epidemiol., vol. 93, p. 102693, 2024.

[8] A. Passaro et al., "Cancer biomarkers: emerging trends and clinical implications for personalized treatment," Cell, vol. 187, no. 7, pp. 1617–1635, 2024.

[9] T. C. Dakal et al., "Emerging methods and techniques for cancer biomarker discovery," Pathol.-Res. Pract., p. 155567, 2024.

[10] R. Ramos et al., "Lung Cancer Therapy: The Role of Personalized Medicine," Cancers, vol. 17, no. 5, p. 725, 2025.

[11] V. D. Ramaswamy and M. Keidar, "Personalized plasma medicine for cancer: transforming treatment strategies with mathematical modeling and machine learning approaches," Appl. Sci., vol. 14, no. 1, p. 355, 2023.

[12] J. van Leuven et al., "Framework for implementing individualised dosing of anti-cancer drugs in routine care: overcoming the logistical challenges," Cancers, vol. 15, no. 13, p. 3293, 2023.

[13] F. Amaro, M. Carvalho, M. de L. Bastos, P. Guedes de Pinho, and J. Pinto, "Pharmacometabolomics applied to personalized medicine in urological cancers," Pharmaceuticals, vol. 15, no. 3, p. 295, 2022.

[14] A. Kumar and D. Metta, "AI-driven precision oncology: predictive biomarker discovery and personalized treatment optimization using genomic data," Int J Adv Res Publ Rev, vol. 1, no. 3, pp. 21–38, 2024.

[15] K. S. Adewole et al., "A systematic review and meta-data analysis of clinical data repositories in Africa and beyond: recent development, challenges, and future directions," Discov. Data, vol. 2, no. 1, p. 8, 2024.

[16] H. Ahmed, S. Hamad, H. A. Shedeed, and A. S. Hussein, "Enhanced deep learning model for personalized cancer treatment," IEEE Access, vol. 10, pp. 106050–106058, 2022.

[17] F. Lococo et al., "Implementation of artificial intelligence in personalized prognostic assessment of lung cancer: a narrative review," Cancers, vol. 16, no. 10, p. 1832, 2024.

[18] S. Herráiz-Gil, E. Nygren-Jiménez, D. N. Acosta-Alonso, C. León, and S. Guerrero-Aspizua, "Artificial Intelligence-Based Methods for Drug Repurposing and Development in Cancer," Appl. Sci., vol. 15, no. 5, p. 2798, 2025.

[19] J. Jian, D. He, S. Gao, X. Tao, and X. Dong, "Pharmacokinetics in pharmacometabolomics: towards personalized medication," Pharmaceuticals, vol. 16, no. 11, p. 1568, 2023.

[20] M. Elhaie, A. Koozari, and D. Shahbazi-Gahrouei, "Machine learning and neural network approaches for enhanced measuring and prediction of radiation doses," J. Radiat. Res. Appl. Sci., vol. 18, no. 1, p. 101252, 2025.

[21] J. Zhang, Y. Lei, J. Xia, M. Chao, and T. Liu, "Federated learning for enhanced dose–volume parameter prediction with decentralized data," Med. Phys., vol. 52, no. 3, pp. 1408–1415, 2025.

[22] S. Pati et al., "Federated learning enables big data for rare cancer boundary detection," Nat. Commun., vol. 13, no. 1, p. 7346, 2022.

[23] K. Cao et al., "A multicenter bladder cancer MRI dataset and baseline evaluation of federated learning in clinical application," Sci. Data, vol. 11, no. 1, p. 1147, 2024.

[24] Samira Alipour, "Genomics of Drug Sensitivity in Cancer (GDSC)." Accessed: Jun. 20, 2025. [Online]. Available: https://www.kaggle.com/datasets/samiraalipour/genomics-of-drug-sensitivity-in-cancer-gdsc

[25] R. Shaik, B. L. Swaroopa, D. Mounika, S. Doddy, B. Yasaswini, and Y. Gottumukkala, "Optimal Drug Dosage Control Strategy System using Decision Tree Regression," Macaw Int. J. Adv. Res. Comput. Sci. Eng., vol. 11, no. 1, pp. 35–45, 2025.

Hybrid Vision Transformer and MLP-Mixer for Epileptic Seizure Detection in Intracranial EEG

Thouraya Guesmi[1], Abir Hadriche[2], Nawel Jmail[3]
Miracl Lab, Sfax University, Tunisia[1, 3]
Gabes National Engineering School, Gabes University, Tunisia[1]
Digital Research Center of Sfax, Tunisia[2, 3]
Regim Lab-ENIS, Sfax University, Tunisia[2]

***Abstract*—Accurate and timely seizure detection is essential for effective epilepsy management, and automated systems can play a valuable role in supporting clinical practice. In this study, we introduce a hybrid approach that uses time-frequency representations of Intracranial electroencephalography (iEEG) signals filtered at High-Frequency Oscillations (HFOs) bands as input to different convolutional neural network (CNN) backbones for feature extraction, followed by classification with either a Vision Transformer (ViT) or MLP-Mixer. This work establishes a systematic, comparative framework for benchmarking hybrid CNN-ViT against CNN-MLP-Mixer, providing a critical new reference for automated epileptic seizure detection within HFOs filtered iEEG signals. Extensive evaluation demonstrates that the ViT consistently achieves superior performance, with an EfficientNetB0-ViT model attaining remarkable accuracy (97.85%) and specificity (98.92%). Crucially, the MLP-Mixer emerges as a highly competitive alternative, exhibiting strong recall capabilities that make it suitable for applications where missing a seizure is not an option. Overall, our findings suggest that self-attention mechanisms in ViTs provide a distinct advantage for capturing complex seizure dynamics, yet MLP-based models present a powerful, efficient option.**

Keywords*—Vision transformer; MLP-Mixer; iEEG; HFOs; ResNet; GoogleNet; EfficientNetB0

I. Introduction

Epilepsy represents a major global health challenge, characterized by recurrent, unprovoked seizures that disrupt neural function and impose a significant burden on the quality of life for over fifty million individuals worldwide [1]. The clinical management of epilepsy, particularly drug-resistant forms, relies heavily on the precise identification and localization of seizure events [2]. Consequently, the development of accurate and reliable seizure detection systems is paramount for enabling timely intervention and improving patient outcomes [3][4].

Intracranial electroencephalography (iEEG) is commonly used to evaluate patients with refractory epilepsy before surgery. It offers detailed recordings from the brain's surface or deeper areas. Analyzing iEEG visually requires a significant amount of time for neurologists, which slows down clinical workflow. Automated detection algorithms that can accurately review long recordings are becoming increasingly popular due to the need for efficiency and precision.

The evolution of these automated systems has progressed from traditional machine learning techniques, which depend on manually engineered features, to deep learning systems that learn discriminative patterns from raw or preprocessed data [5]. Among these, the ability to analyze time-frequency representations of iEEG signals has been demonstrated by Convolutional Neural Networks (CNNs). They demonstrated exceptional proficiency in analyzing time-frequency representations of iEEG signals, effectively identifying localized morphological patterns associated with seizure onset[6]. Despite their strengths, a fundamental constraint of CNNs is their limited receptive field, which can hinder their ability to model the long-range spatiotemporal dependencies that characterize the dynamic propagation of seizures through neural networks [7].

To address this limitation, Vision Transformers (ViT) have the ability to capture global contextual relationships across an entire input sequence through their self-attention mechanism, potentially offering a more nuanced understanding of complex ictal dynamics [8] [9].

In recent research, hybrid architectures have been explored that combine the local feature extraction capabilities of CNNs with the global contextual modeling of ViTs, which have resulted in significant gains in both accuracy and generalizability for seizure detection tasks.[10].

Parallel to these developments, the MLP-Mixer architecture presents a notably different approach. As articulated by Tolstikhin et al. (2021), "In contrast to ViT, which relies on self-attention, the MLP-Mixer proposes an architecture based solely on Multi-Layer Perceptrons (MLPs) for spatial and channel mixing, offering a competitive accuracy-efficiency trade-off[11]". While MLP-Mixers have shown promise in other domains, their application to the specific challenge of detecting epileptic seizures from iEEG signals filtered in the high-frequency oscillation (HFO) bands remains a critical, largely unexplored research gap.

This study addresses this gap by investigating the fundamental trade-offs between these novel sequence modeling approaches when applied to a highly sensitive biomarker: HFO-filtered iEEG data. This leads to our core research question: What are the distinct performance, efficiency, and mechanism trade-offs between hybrid CNN-ViT and CNN-MLP-Mixer systematically benchmarked for automated epileptic seizure detection in the HFOs domain?

In summary, our contributions are as follows: We introduce a systematic and novel comparative framework for benchmarking the performance and mechanism trade-offs of hybrid CNN-Vision Transformer and CNN-MLP-Mixer architectures.

We establish a novel deep learning reference for the automated detection of epileptic seizures in the High-Frequency Oscillation (HFO) domain of iEEG, addressing a major gap in the current literature.

We provide a detailed performance analysis that contrasts the advantages of the self-attention mechanism versus the purely algebraic mixing of MLP-Mixer, highlighting critical considerations for clinical deployment.

We systematically evaluate six widely used CNN backbones (VGG19, ResNet18, ResNet50, ResNet101, GoogleNet, EfficientNet-B0) to provide a comprehensive baseline for feature extraction in HFO analysis.

This study is organized as follows: Following the introduction, a theoretical framework reviews relevant literature (Section II). Details on materials and methods are provided in Section III. The results are presented in Section IV, followed by a comprehensive discussion of the findings, their implications, and the limitations of the study in Section V. The study concludes in Section VI.

II. Related Works

The pursuit of automated seizure detection has progressively shifted from reliance on manually crafted features to the use of deep learning models. Within this evolution, [12] showed that HFOs detected by neuromorphic neural networks are strong predictors of seizure prognosis across various recording modalities, including iEEG. This work demonstrates the clinical relevance of HFO-based detection but relies on hand-crafted feature engineering rather than end-to-end deep learning. Other research works [13][14] substantiate the strong correlation between HFOs and epileptogenic tissue, establishing a solid foundation for their use in automated prognosis systems. The strength of these studies lies in validating HFOs as critical biomarkers for seizure detection. However, a key weakness is the limited exploration of modern sequence modeling architectures (Transformers, MLP-Mixers) that could better capture the temporal dynamics of HFO patterns. The undesirable assumption often made is that traditional feature engineering is sufficient to capture the intricate, non-linear dynamics of HFOs.

A critical step in applying modern deep learning architectures to signal data involves transforming one-dimensional time series into a format that can be used with image-based models. In this context, Zhuohan Wang et al [15] demonstrated that transforming iEEG signals into scalograms facilitates the use of image-based deep learning models. This approach enables the model to simultaneously leverage temporal and spectral information. We build upon this foundation, using time-frequency transformation as the essential input for our Vision Transformer and MLP-Mixer models.

In the context of iEEG analysis, CNNs have demonstrated superior performance in detecting epileptic seizures by learning local discriminative patterns in spectrograms, with variants such as ResNet, GoogLeNet and VGG being widely used to extract meaningful features from biomedical signals [16]. While excellent at capturing features within a limited spatial context, they often struggle to integrate information across distant regions of a scalogram, which is crucial for understanding the full spatial-temporal propagation of a seizure. This limited and fixed receptive field is the core weakness of CNN-only approaches, limiting their ability to model global seizure connectivity. This inadequacy is precisely why we propose a hybrid framework, pairing the local robustness of CNNs with sequence models capable of global context aggregation.

The MLP-Mixer architecture challenges the dominance of convolutions and self-attention by relying solely on a multilayer perceptron, applied separately to spatial (token-mixing) and channel (feature-mixing) dimensions [11]. Its simplicity and computational efficiency make it attractive for EEG-based applications. Recent studies have shown that MLP-Mixers, enhanced with attention mechanisms, can effectively decode EEG motor imagery tasks [17]. Despite this, the specific efficacy of the MLP-Mixer for the nuanced task of iEEG-based seizure detection, particularly when operating on HFOs, remains an open question. The current literature lacks a systematic comparison between the self-attention mechanism (ViT) and the purely algebraic mixing (MLP-Mixer) to determine which is inherently superior for modeling HFO dynamics. This forms the central motivation and unique contribution of our study.

III. Materials and Methods

This section details the methodology adopted, from the preparation of raw data to the configuration of hybrid models. Our study proposes a systematic comparative framework for hybrid deep learning CNN-VIT and CNN-MLP-Mixer applied to epileptic seizure detection in iEEG filtered in HFO bands. This study proposes a hybrid pipeline for binary classification between ictal and interictal states. The core approach involves using Convolutional Neural Networks (CNNs) backbones (EfficientNetB0, ResNet18, ResNet50, ResNet101, GoogleNet and VGG19) to extract features from time-frequency images of iEEG signals, which are then classified by either a Vision Transformer (ViT) or an MLP-Mixer. The overall workflow is depicted in Fig. 1.

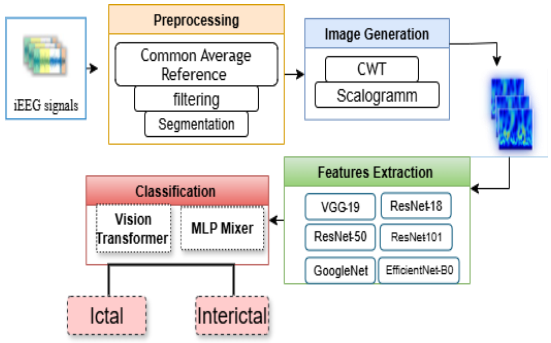


Fig. 1. Workflow of the proposed method.

A. Data Description

This study utilized a subset of the publicly available HUP iEEG dataset [18], which comprises intracranial EEG recordings from patients undergoing evaluation for drug-resistant epilepsy at the Hospital of the University of Pennsylvania. From the original cohort of 58 subjects, we selected 21 patients whose recordings were sampled at 1024 Hz, a frequency sufficient for the reliable analysis of High-Frequency Oscillations (HFOs). The dataset includes ictal and interictal recordings, to precise annotations marking the beginning (onset) and end (offset) of seizure events, allowing for clear identification of ictal periods.

B. Data Processing

The iEEG data preprocessing pipeline consisted of three main stages: segmentation, filtering/re-referencing, and time-frequency transformation.

Segmentation and labeling were performed for the binary classification task of distinguishing ictal from interictal states. The continuous iEEG recordings were segmented into 1-second epochs. Ictal segments were extracted from annotated seizure intervals (onset to offset). Each segment was assigned a binary label (ictal=1, interictal=0), defining the classification target as $y \in \{0,1\}$.

Channels marked as artifactual in the original dataset metadata were discarded. The signals were then re-referenced using a Common Average Reference (CAR) filter computed from all remaining valid channels, a crucial step for reducing common-mode noise. Subsequently, a zero-phase finite impulse response (FIR) filter was applied to isolate the High-Frequency Oscillation (HFO) band of interest (80-250 Hz).

To leverage image-based deep learning models, each 1-second HFO-filtered segment was converted into a 2D time-frequency (TF) (see Fig. 2) representation using the Continuous Wavelet Transform (CWT) with a complex Morlet wavelet. The wavelet function is defined as:

$$\varphi(t) = C_\sigma \pi^{-1/4} e^{-t^2/2} e^{i\sigma t} \quad (1)$$

where, C_σ is the normalization constant and σ is the central frequency parameter. The 2D representation results were scaled to dimensions of 224 × 224 × 3 pixels as input images for the proposed model. The final dataset comprised 4092 samples, which were partitioned into training (80%), validation (10%), and test (10%) sets.

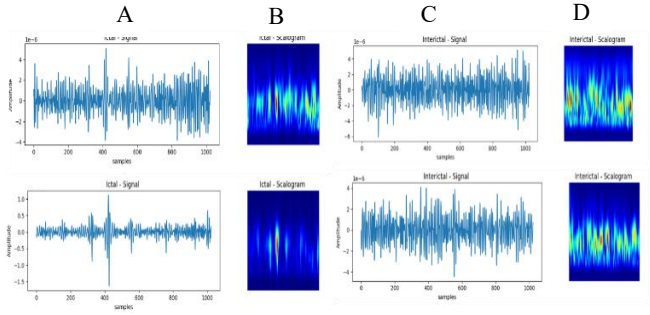


Fig. 2. Transformation of signal into 2D representation: (A) ictal representation in time, (B) ictal representation in TF, (C) interictal representation in time, (D) interictal representation in TF.

C. Model Development

1) CNN backbones: Resnet (Residual Network): Resnet is a fundamental contribution to deep learning, specifically addressing the problem of performance degradation when training extremely deep convolutional neural networks (CNNs). Its main innovation lies in the introduction of residual blocks, which enable the learning of residual mapping functions via skip connections. In the ResNet architecture, the fundamental residual connection in ResNet consists of summing the original input x with the learned transformation F(x) to produce the block's output, according to the following equation:

$$y = F(xw_i) + x \quad (2)$$

ResNet50: This architecture comprises 50 layers, and its structure incorporates batch normalization layers and ReLU activations to stabilize training. Features are extracted from the avgpool layer, resulting in a 2048-dimensional vector.

ResNet18: With 18 layers, its skip connections between convolutional layers, effectively solve the vanishing gradient problem. Although less deep.

ResNet101: This deep architecture has 101 layers and optimizes residual learning between the inputs and outputs of convolutional blocks, enabling richer feature extraction.

VGG19: It is characterized by its fixed depth of 19 layers and its uniform architecture using exclusively 3×3 convolutions.

GoogleNet (Inception): It has a 22-layer architecture optimized for multi-scale feature extraction via its Inception modules, which incorporate Inception blocks that use parallel multi-kernel convolutions. This design enables the model to identify patterns across various spatial scales.

EfficientNet-B0: It is an architecture optimized via compound scaling that balances depth, width, and resolution. The model uses MBConv blocks with expansion-reduction mechanisms and channel attention (SE modules).

For each CNN, the fully connected classification layer is removed. Consequently, the feature extraction configuration for each model is detailed in Table I, which shows the dimensions of the final feature maps extracted from the CNN backbones for an input image size of 224×224×3.

TABLE I. DIMENSIONS OF THE FINAL FEATURE MAPS OF THE CNN BACKBONES

Backbone	Feature Map [B, C, H, W]
ResNet18	[B, 512, 7, 7]
ResNet50	[B, 2048, 7, 7]
ResNet101	[B, 2048, 7, 7]
VGG19	[B, 512, 7, 7]
GoogleNet	[B, 1024, 7, 7]
EffecientNet-B0	[B, 1280, 7, 7]

B = batch size.

C = number of channels (feature dimension).

H, W = spatial dimensions (here 7×7 for input 224×224).

2) Vision transformer: Transformers are an important advancement in deep learning. They were first created for natural language processing [19] but now help solve many sequence modeling problems. Because they can handle complex, long-range relationships, they are useful in areas like computer vision [19] and biomedical data analysis [20]. Specifically, The Vision Transformer (ViT) is a transformer that has been adapted for image inputs. It works by dividing images into patches, which are then embedded linearly. Each patch vector is then enriched with a position embedding to preserve spatial information. The embeddings are then subjected to a series of Transformer blocks. The blocks that make it up include Multi-Head Self-Attention (MHSA), Layer Normalization (LN), and Feed-Forward Network (FFN). The Multi-Head Self-Attention (MSA) calculates the interactions between all tokens in the sequence. For each head, linear projections generate the matrices query (Q), key (K), and value (V) vectors, as shown in the following equations:

$$\mathrm{Q}=\mathrm{XW}^{\mathrm{Q}}\ ,\quad K=XW^{K}\ ,\ V=XW^{V} \tag{3}$$

Let W^{Q}, W^{K}, and W^{V} denote the parameter weight matrices. Then, the attention mechanism computes a weighted sum of the values, where attention coefficients are derived from a normalized query-key similarity as represented in the following equation:

$$Attention(Q,K,V)=\mathrm{softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{4}$$

in which, normalization by $\sqrt{d_k}$ aims to stabilize gradients during learning. Subsequently, the output of the attention layer was subjected to Layer Normalization (LN), a process that has been demonstrated to stabilize the training process (see equation below).

$$LN(z)=\frac{z-\mu}{\sqrt{\sigma^{2}+\epsilon}}.\gamma+\beta \tag{5}$$

The input's meaning and standard deviation are μ and , while $\gamma+\beta$ are acquired parameters. A non-linear GELU activation separates two linear layers in a Feed-Forward Network (FFN):

$$FFN(z)=\mathrm{W}_{2}.GELU(\mathrm{W}_{1}.z+\mathrm{b}_{1})+\mathrm{b}_{2} \tag{6}$$

With the b being biases and W representing weight matrices.

3) Proposed CNN-Vision Transformer hybrid model: This proposed hybrid model combines the strength of CNNs in extracting features, which are subsequently converted into tokens and processed by a Transformer encoder. This fusion skillfully leverages CNNs for capturing detailed local features and the Transformer for global context modeling via the self-attention mechanism, the architecture proceeds as follows:

Feature extraction by CNN: Let an input image (spectrogram) be denoted as $I\in\mathbb{R}^{224\times224\times3}$ which is processed by a pre-trained CNN network(ReNet18, ResNet50, ResNet101, VGG19, GoogleNet, EfficientNet-B0), producing a feature map $F\in\mathbb{R}^{7\times7\times C_{nn}}$ whose final convolutional outputs are 7×7 spatial maps for 224×224 inputs, the output of each CNN is detailed in Table I where C_{nn} represents the number of output channels.

The transition to the ViT: The resulting feature maps for each CNN are linearly projected to embedding dimension $D=768$ using 2D (1×1) conventional with (padding='same') followed by reshaping to obtain $(7\times7,768)$. Consequently, $N=7\times7=49$ tokens. This embedding dimension (D=768) is adopted from the ViT-Base specification, which balances model capacity with computational efficiency for sequences of moderate length. Then, a learnable [CLS] token is prepended to the token sequence. Learnable positional embedding $E=49+1=50$ is added to preserve spatial information.

Transformer Encoder: processes the input sequence through twelve encoder layers, where each layer helps the model understand context and share information between tokens. The model can focus on multiple parts of the input sequence simultaneously by using a multi-head self-attention mechanism in each encoder layer that has twelve attention heads.

Classification: Each encoder layer also has an MLP block with 3072 dimensions (3072 = 4×768 to provide sufficient non-linear transformation capacity) and uses GELU activation. Following the encoder layers, the final state of the classification token is extracted, layer-normalized, and processed by a multilayer perceptron (MLP) classification head to produce the final class predictions, where the classification token gathers information from all tokens using self-attention. The final prediction is obtained through a sigmoid activation function:

$$MLP-Head(z_{cls})=W_{2}.GELU(W_{1}.LayerNorm(z_{cls})+b_{1})+b_{2} \tag{7}$$

$$\hat{y}=\sigma(W_{class}.MLP-Head(z_{cls})+b_{class}) \tag{8}$$

where, $z_{cls}\in\mathbb{R}^{768}$ is the final cls token, $W_{1}\in\mathbb{R}^{3072\times768}$, $W_{1}\in\mathbb{R}^{768\times3072}$ and σ is the sigmoid function for binary classification. We use Transformer-related parameters including an embedding dimension of 768, 12 encoder layers, 12 attention heads, and an MLP hidden size of 3072 following the standard ViT-Base configuration. This ensures consistency with well-established architecture and allows our hybrid model to benefit from proven practices for effective global context modeling. The complete architecture is illustrated in Fig. 3 and the detailed algorithm for the model is given below (see Algorithm 1).

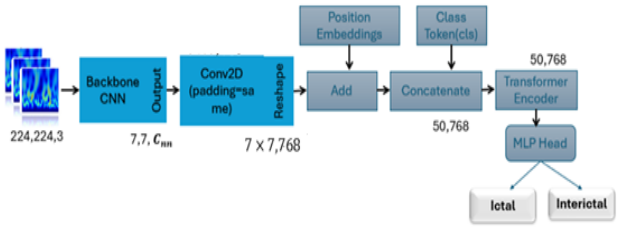


Fig. 3. Architecture of CNN vision transformer.

Algorithm 1 : Hybrid CNN-Vision Transformer

Input:
Time frequency (TF) image $I \in \mathbb{R}^{224\times224\times3}$

Target $y_i \in$ [0 1] 1=ictal; 0=interictal

Output:

Predicted class probability $\hat{y} \in$ [0 1]

Begin:
Data selection

IEEG data

Pre-processing

filtering/re-referencing

Segmentation

2D transformation using CWT

Resize image to 224×224

Labelling $y = \begin{cases} 1: ictal \\ 0: interictal \end{cases}$

Step1: CNN Features Extraction

F_cnn = CNN(I) $F \in \mathbb{R}^{7\times7\times C_{nn}}$

Step2 Token Projection

Conv2D (1 × 1) (padding= same); (7,7,768)

Reshape (7*7,768)

Z_flat = Reshape (Conv2D(F_cnn)); Z_flat $\in \mathbb{R}^{49\times768}$

Z = Concat([CLS], Z_flat) + Positional Embedding

$Z \in \mathbb{R}^{50\times768}$

Step3: Transformer Encoder
for layer = 1 to 12:

-Head Self-Attention

Z_norm = Layer Norm(Z)

$Attention(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

Z_att = MultiHead (Attention) + Z

-Forward Network

$$FFN(z) = W_2.GELU(W_1.z + b_1) + b_2$$

Step4: MLP Classification Head

$MLP - Head(z_{cls}) = W_2.GELU\,(W_1.LayerNorm(z_{cls}b_1) + b_2$

$\hat{y} = \sigma(W_{class}.MLP - Head(z_{cls}) + b_{class})$

Trainnig

Loss= binary cross entropy

Train model model.fit(X_train, y_train)

Evaluation

model.evaluate(X_test, y_test)

End

4) Proposed CNN MLP-Mixer hybrid model: The MLP-Mixer approach eliminates both convolution and attention operations, in contrast to the transformer architecture, by employing multilayer perception (MLPs) organized into token-mixing and channel-mixing layers. This architecture offers a conceptually simpler alternative to self-attention while maintaining competitive performance through pure algebraic mixing operations [11].

For an input matrix $X \in \mathbb{R}^{S\times D}$, where S is the number of tokens and D is the number of feature dimensions; the fundamental operations of the MLP-Mixer are defined as follows:

Token mixing (modelling spatial relationships):

$$U_{*,i} = X_{*,i} + W_2\sigma(W_1 LayerNorm(X)_{*,i}) \quad (9)$$

for $i = 1,..D$

σ is the GELU activation function.

$X_{*,i}$: all tokens for a channel i (the i-th column of X).

W : Weight of MLP mixing tokens (for $W_1 \in \mathbb{R}^{S'\times S}$, $W_2 \in \mathbb{R}^{S\times S'}$ with S' being the intermediate dimension.

Channel mixing (modelling relationships between characteristics):

$$Y_{j,*} = U_{j,*} + W_4\sigma(W_3 LayerNorm(U)_{j,*}) \quad (10)$$

for $j = 1, \ldots., S$

$U_{*,i}$: all tokens for channel j (the jth column of U).

W : Weights of the MLPs of channel mixing (for $W_3 \in \mathbb{R}^{D'\times D}$, $W_4 \in \mathbb{R}^{D\times D'}$ with D' being the intermediate dimension).LayerNorm refers to layer normalization.

For the final classification a Global average pooling is applied followed by a linear classifier:

$$\hat{y} = \sigma(W_{class}.GlobalAvreagePooling(Y) + b_{class}) \quad (11)$$

Architectural Configuration: (The feature extraction step is described in the previous section), to configure the model architecture, the following approach was adopted a hidden dimension of D= 192, which aligns with Base-type models and strikes a balance between expressiveness and computational limits [11]. Eight layers (L=8) are stacked to reduce overfitting and handle dependencies in sequences of moderate length.

The architecture features an intentional imbalance in its MLP dimensions. For processing spatial information (token-mixing), a bottleneck of 96 units is employed to promote efficiency. For transforming feature information (channel-mixing), a wide network of 768 units is used to facilitate richer representations. This asymmetric design reflects the observation that channel-wise feature interactions typically require higher capacity than spatial token interactions for time-frequency representations. The detailed algorithm for the model is given in Algorithm 2, and the architecture is illustrated in Fig. 4.

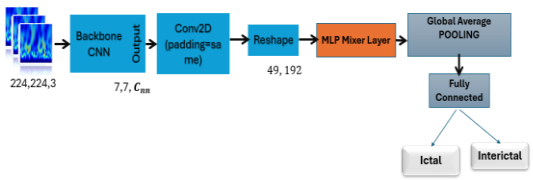


Fig. 4. Architecture of CNN MLP-Mixer.

Algorithm 2: Hybrid CNN-MLP-Mixer

Input:

Time frequency (TF) image $I \in \mathbb{R}^{224\times224\times3}$

Target $y_i \in$ [0 1] 1=ictal; 0=interictal

Output:

Predicted class probability $\hat{y} \in$ [0 1]

Begin:

Data selection

IEEG data

Pre-processing

filtering/re-referencing

Segmentation

2D transformation using CWT

Resize image to 224×224

Labelling $y = \begin{cases} 1: ictal \\ 0: interictal \end{cases}$

Step1: CNN Features Extraction

F_cnn = CNN(I) $F \in \mathbb{R}^{7\times7\times C_{nn}}$

Step2 Features Projection

Conv2D (1 × 1) (padding= same); (7,7,192)

Reshape (7*7,192)

Z = Reshape (Conv2D(F_cnn)); $Z \in \mathbb{R}^{49\times192}$

Step 3: MLP-Mixer Layers (8 blocks)

Token-Mixing MLP (spatial)

U_norm = LayerNorm(Z)

for i = 1 to 192: Per-channel processing

$$U_{*,i} = X_{*,i} + W_{2^\sigma}(W_1 LayerNorm(X)_{*,i})$$

$W_1 \in \mathbb{R}^{96\times49}$ $W_3 \in \mathbb{R}^{49\times96}$

for j= 1 to 49: Per-token processing

$$Y_{j,*} = U_{j,*} + W_{4^\sigma}(W_3 LayerNorm(U)_{j,*})$$

$W_3 \in \mathbb{R}^{768\times192}$ $W_4 \in \mathbb{R}^{192\times768}$

Step 4: Classification

x_pool = GlobalAveragePooling(Y) ; x_pool $\in \mathbb{R}^{192}$

$\hat{y} = \sigma(\text{x_pool} \times W_{class}. + b_{class})$ $W_{class} \in \mathbb{R}^{(1,192)}$

Trainnig

Loss= binary cross entropy

Train model model.fit(X_train, y_train)

Evaluation

model.evaluate(X_test, y_test)

End

5) Simulation Map: All models were implemented in Python 3.9 using PyTorch 2.0.1. Experiments were conducted on Google Colab using GPU acceleration. The following Table II shows the different hyperparameters used in the study.

TABLE II. CONFIGURATION OF TRAINING HYPERPARAMETERS

Hyperparameters	Value
Optimizer	ADAM
Loss function	Binary cross entropy (BCE)
Batch Size	32
Epochs	80(early stopping with patience=10
Learning Rate	1e-4
Dropout	0.1

IV. RESULTS

The models were evaluated using standard binary classification metrics: accuracy, precision, recall (sensitivity), specificity, F1-score (see Table III) and the area under the receiver operating characteristic curve (AUC-ROC). These metrics were derived from the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in the predictions.

TABLE III. EVALUATION METRICS

Metrics	Equations
Accuracy(ACC)	$\frac{TP+TN}{TP+FP+FN+TN}$
Precision (Pr)	$\frac{TP}{TP+FP}$
Recall (Se)	$E = \frac{TP}{TP+FN}$
Specificity (Sp)	$1 - \frac{FP}{FP+TN}$
F1-score	$F1-score = \frac{2\times TP}{(2\times TP+FP+FN)}$

Table IV and Table V present the comprehensive performance evaluation of Vision Transformer (ViT) and MLP-Mixer classifiers across six CNN backbones. The results reveal a consistent performance hierarchy, with the EfficientNetB0-ViT configuration achieving the highest overall scores (Accuracy: 97.85%, Specificity: 98.92%). GoogleNet was also consistent, performing well with both classifiers (96.45% with ViT and 95.20% with MLP-Mixer) and stood out in recall metrics. However, the results obtained with VGG19 are the least optimal.

TABLE IV. PERFORMANCE COMPARISON OF CNN-VISION TRANSFORMER

Model	ACC%	Pr %	Se%	Sp%	F1-score %
EfficientNETB0-ViT	**97.85**	98.32	**96.88**	**98.92**	97.59
ResNet18-ViT	95.35	94.90	92.70	95.95	93.78
ResNet50-ViT	94.80	95.42	93.25	96.35	94.32
ResNet101 ViT	95.25	95.05	95.50	95.00	95.27
GoogleNet VIT	96.45	97.15	95.20	97.85	96.16
VGG19 VIT	93.20	93.65	91.85	94.70	92.74

TABLE V. PERFORMANCE COMPARISON OF CNN MLP-MIXER

Model	ACC%	Pr %	Se%	Sp%	F1-score %
EfficientNETB0MLP-Mixer	97.08	96.55	94.45	97.35	95.48
ResNet18 MLP-Mixer	92.90	91.35	94.25	91.55	92.78
ResNet50- MLP-Mixer	93.25	93.85	91.55	95.05	92.68
ResNet101 MLP-Mixer	93.80	92.45	95.10	92.50	93.76
GoogleNet MLP-Mixer	95.20	94.80	96.75	93.65	95.76
VGG19 MLP-Mixer	91.75	92.20	90.15	93.55	91.16

This comparative analysis provides a foundation for further investigation of operational characteristics, such as ROC curves. As illustrated in Fig. 5, the comparative ROC curves demonstrate the discriminatory power of the various models. The EfficientB0-ViT model demonstrates a remarkable performance with an Area Under the Curve (AUC) of 0.979, the subsequent models are GoogleNet-ViT (AUC = 0.968) and EfficientB0-MLP (AUC = 0.963). The ViT curves have been shown to demonstrate a systematic superiority that confirms and substantiates its efficacy in differentiating between ictal and interictal states.

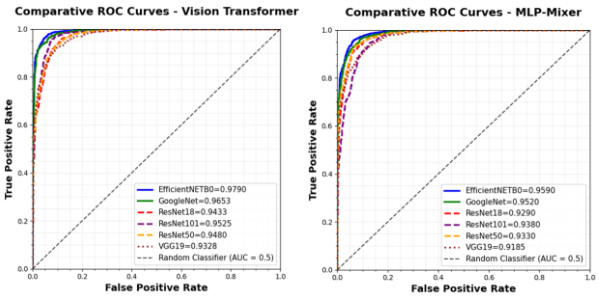


Fig. 5. Comparative ROC Curve: (A) Curve Roc VIT, (B) Curve ROC MLP-mixer.

To understand the impact of the model's performance, we analyzed normalized confusion matrices. These matrices, shown in Fig. 6, reveal each architecture's specific error profile, particularly their trade-offs between false positives and false negatives in distinguishing ictal from interictal states.

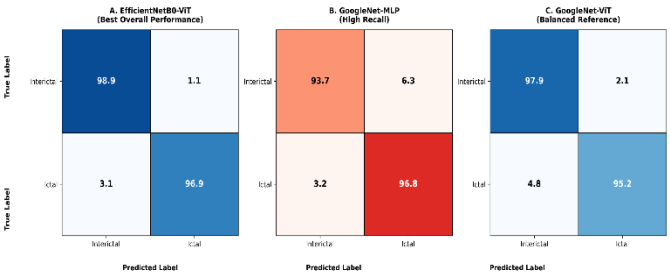


Fig. 6. Normalized confusion matrices for the three best-performing models: (A) EfficientNetB0-ViT, (B) GoogleNet-MLP, (C) GoogleNet-VIT.

A detailed analysis of the error profiles, provided by the confusion matrices in Fig. 6, offers crucial insights for clinical application. The primary strength of EfficientB0-ViT [see Fig. 6(A)] shows only 3.1% false negatives and 1.1% false positives. This reliability makes it ideal for clinical deployment, where the cost of a false alarm (false positive). Interestingly, the GoogleNet-MLP Mixer [see Fig. 6(B)] model has the best recall rate (96.8%). However, it has a higher rate of false positives (6.3%) compared to false negatives (3.2%), which is important for minimizing false negatives (missed seizures). Meanwhile, GoogleNet-ViT [see Fig. 6(C)] offers balanced performance with moderate error rates (false negatives at 4.8% and false positives at 2.1%).

Fig. 7 provides a direct comparative analysis of the performance differential between Vision Transformer and MLP-Mixer architectures. The histogram quantifies the percentage point advantage of ViT over MLP-Mixer for the EfficientNetB0 backbone.

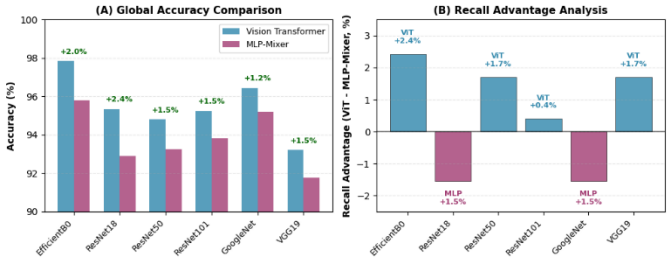


Fig. 7. Performance gain.

Fig. 7 presents the performance differential between Vision Transformer and MLP-Mixer architectures when paired with the EfficientNetB0 backbone. The results demonstrate a consistent advantage for the ViT model across all evaluation metrics, with particularly notable margins in recall and accuracy. However, the MLP-Mixer maintains competitive performance, trailing by less than 2.5 % points in all categories. This narrow performance gap, especially in specificity where both architectures excel, underscores the MLP-Mixer's viability as an efficient alternative to attention-based models for seizure detection tasks.

The findings confirm that while the self-attention mechanism in ViT yields superior overall performance, the MLP-Mixer achieves remarkably close results. This suggests context-dependent applicability: ViT for maximum diagnostic precision versus MLP-Mixer for scenarios prioritizing computational efficiency alongside strong detection capability.

The superior performance of the proposed EfficientNetB0-ViT model is evident when benchmarked against recent literature, as summarized in Table VI. Our method establishes a new accuracy of 97.9%, representing a significant improvement over previous approaches for similar seizure detection challenges.

TABLE VI. COMPARISON OF PREVIOUS WORKS

Authors	Year	Methods	Accuracy%
S.Roy et al.[21]	2019	ChronoNet	90.60
RaviPrakash et al.[22]	2020	1D-CNN-LTSM	89.73
Priya et al.[23]	2022	ResNet50	82.80
Rukshar et al. [24]	2023	Vision Transformer	89.07
Gupta et al.[25]	2024	FUPTBSVM(hybrid model)	88.66
Qi et al.[26]	2025	Vision Transformer	93.65
OURS		**EfficientNet-B0 ViT**	**97.9**

V. DISCUSSION

The experimental results provide a clear answer to the primary research question: Vision Transformer (ViT) classifiers consistently outperform MLP-Mixer counterparts when integrated with CNN backbones for iEEG-based seizure detection. The superior performance of the EfficientNetB0-ViT model (97.85% accuracy, 98.92% specificity) establishes the effectiveness of self-attention mechanisms for capturing the spatiotemporal dynamics of epileptic seizures. This advantage can be theoretically explained by the ViT's ability to model global dependencies across the full time–frequency representation, a capability that aligns well with the distributed neural networks underlying seizure propagation. The observed performance gap between ViT and MLP-Mixer models provides valuable insight into the architectural requirements for effective seizure detection. The consistent advantage of ViT across multiple metrics suggests that the ability to dynamically weight different regions of the time-frequency representation (a core property of self-attention mechanisms) provides significant value for this task. This is particularly relevant for seizures that manifest as evolving patterns across both time and frequency domains. Our results for the MLP-Mixer architecture align with findings from computer vision (Tolstikhin et al., 2021), where MLP-based models have demonstrated competitive performance despite their conceptual simplicity. Extending this observation to biomedical signal analysis, our findings demonstrate that MLP-Mixers constitute a viable alternative to attention-based architectures for time-series classification, offering a balance between interpretability and efficiency. The strong correlation between CNN backbone quality and final performance (evidenced by EfficientNetB0's superior results across both classifiers) underscores the continued importance of effective feature extraction. Indeed, the identification of dominant patterns is critical for characterizing neurological disorders [27]. Even highly expressive classification heads such as ViT cannot compensate for suboptimal convolutional representations, confirming the hierarchical dependency between feature extraction and global context modeling in hybrid deep-learning architectures.

VI. CONCLUSION

This study presented a systematic and novel comparative framework for the automated detection of epileptic seizures from intracranial EEG (iEEG) signals filtered in the High-Frequency Oscillation (HFO) band. By integrating convolutional feature extractors with advanced sequence-modeling architectures Vision Transformer (ViT) and MLP-Mixer we established a comprehensive evaluation of hybrid deep-learning approaches for seizure detection.

The proposed benchmarking protocol represents a novel and systematic comparison between CNN-ViT and CNN-MLP-Mixer models specifically tailored to HFO-based iEEG analysis. Across multiple CNN backbones, the ViT consistently outperformed the Mixer counterpart, with the EfficientNetB0-ViT achieving the highest overall accuracy of 97.85 % and specificity of 98.92 %. This performance highlights the ability of self-attention mechanisms to capture the complex spatiotemporal dynamics of seizure propagation. The MLP-Mixer, on the other hand, demonstrated strong recall (up to 96.75 % with GoogleNet-MLP-Mixer), confirming its suitability for scenarios where sensitivity is clinically prioritized, such as continuous monitoring or early-warning systems.

The comparative trends observed in this work have direct practical value: ViT-based hybrids are better suited for high-specificity diagnostic contexts, such as pre-surgical assessment where false positives must be minimized, whereas Mixer-based hybrids provide a high-recall solution ideal for long-term monitoring. These insights contribute to a deeper understanding of how architectural design choices in hybrid deep learning can be aligned with distinct clinical objectives. Moreover, the consistent superiority of ViT across architectures suggests that self-attention represents a robust modeling paradigm for the non-linear, high-frequency dynamics of epileptic activity extending its relevance beyond computer vision to neurophysiological signal analysis.

While the proposed framework demonstrates strong performance, its generalizability should be validated on larger, multi-center datasets including diverse patient populations. In addition, the focus on high-frequency oscillations, though biologically motivated, may have excluded complementary low-frequency information (e.g., delta or theta bands) relevant to seizure onset and propagation. Addressing these limitations will strengthen the translational impact of the approach.

Future research will extend the current binary framework to multi-class classification, encompassing different HFO subtypes and pre-ictal phases. Another direction will involve integrating explainability techniques such as Grad-CAM or LIME to visualize discriminative regions in the time-frequency maps, thereby enhancing clinical interpretability and trust. Finally, incorporating multimodal spectral inputs and transfer-learning strategies will further improve robustness and pave the way for deployment in real-world clinical environments.

ACKNOWLEDGMENT

The authors are grateful to the investigators who shared their iEEG data.

REFERENCES

[1] N. Jmail et al., "Separation between spikes and oscillation by stationary wavelet transform implemented on an embedded architecture," Journal of the Neurological Sciences, Volume 381, 542.2017. doi: 10.1016/j.jns.2017.08.3735.

[2] A. Hadriche, I. Behy, A. Necibi, A. Kachouri, C. Ben Amar, and N. Jmail, "Assessment of Effective Network Connectivity among MEG None Contaminated Epileptic Transitory Events," Comput. Math. Methods Med., vol. 2021, 2021, doi: 10.1155/2021/6406362.

[3] Hadriche, I. ElBehy, A. Hajjej, and N. Jmail, "Evaluation of Techniques for Predicting a Build Up of a Seizure," Lect. Notes Networks Syst., vol. 418 LNNS, no. 1, pp. 816–827, 2022, doi: 10.1007/978-3-030-96308-8_76.

[4] Abir Hadriche,Nawel Jmail "Clinical Images and Medical Case Reports A build up of seizure prediction and detection Software : A review, J Clin Images Med Case Rep. 2021, 2 (2):1087.

[5] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review.," J. Neural Eng., vol. 16, no. 5, p. 51001, Aug. 2019, doi: 10.1088/1741-2552/ab260c.

[6] R. Hussein, H. Palangi, R. Ward, and Z. J. Wang, "Epileptic Seizure Detection: A Deep Learning Approach," pp. 1–12, 2018, [Online]. Available: http://arxiv.org/abs/1803.09848

[7] X. Wu, Z. Yang, T. Zhang, L. Zhang, and L. Qiao, "An end-to-end seizure prediction approach using long short-term memory network," Front. Hum. Neurosci., vol. 17, no. May, pp. 1–10, 2023, doi: 10.3389/fnhum.2023.1187794.

[8] K. Mohiuddin et al., "Retention Is All You Need," Int. Conf. Inf. Knowl. Manag. Proc., no. Nips, pp. 4752–4758, 2023, doi: 10.1145/3583780.3615497.

[9] S. Yuan, K. Yan, S. Wang, J. X. Liu, and J. Wang, "EEG-Based Seizure Prediction Using Hybrid DenseNet–ViT Network with Attention Fusion," Brain Sci., vol. 14, no. 8, 2024, doi: 10.3390/brainsci14080839.

[10] Q. Li, W. Cao, and A. Zhang, "Multi-stream feature fusion of vision transformer and CNN for precise epileptic seizure detection from EEG signals," J. Transl. Med., vol. 23, no. 1, 2025, doi: 10.1186/s12967-025-06862-z.

[11] Tolstikhin et al., "MLP-Mixer: An all-MLP Architecture for Vision," Adv. Neural Inf. Process. Syst., vol. 29, pp. 24261–24272, 2021.

[12] Burelo, M. Sharifshazileh, G. Indiveri, and J. Sarnthein, "Automatic Detection of High-Frequency Oscillations With Neuromorphic Spiking Neural Networks," Front. Neurosci., vol. 16, no. June, pp. 1–17, 2022, doi: 10.3389/fnins.2022.861480.

[13] T. Guesmi, A. Hadriche, and N. Jmail, "Effective connectivity of high-frequency oscillations (HFOs) using different source localization techniques," ISDA 2022, doi: 10.1007/978-3-031-35507-3_36.

[14] T. Guesmi, A. Hadriche, N. Jmail, and C. Ben Amar, "Evaluation of Stationary Wavelet Transforms in Reconstruction of Pure High Frequency Oscillations (HFOs)," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12157 LNCS, no. 80200, pp. 357–363, 2020, doi: 10.1007/978-3-030-51517-1_32.

[15] Z. Wang et al., "EEG-Based Seizure Detection Using Dual-Branch CNN-ViT Network Integrating Phase and Power Spectrograms," Brain Sci., vol. 15, no. 5, pp. 1–24, 2025, doi: 10.3390/brainsci15050509.

[16] Zayneb Sadek, "An Efficient CNN and RNN Hybrid Model for the Detection of Epileptic Seizures in EEG Signals," in IEEE Symposium on Bioinformatics and Bioengineering (BIBE), 2024.

[17] M. Bashar, O. Monjur, S. Islam, M. G. Shams, and N. Quader, "Exploring Synergistic Ensemble Learning: Uniting CNNs, MLP-Mixers, and Vision Transformers to Enhance Image Classification," 2025, [Online]. Available: http://arxiv.org/abs/2504.09076.

[18] and B. L. John M. Bernabei, Adam Li, Andrew Y. Revell, Rachel J. Smith, Kristin M. Gunnarsdottir, Ian Z. Ong, Kathryn A. Davis, Nishant Sinha, Sridevi Sarma, "HUP iEEG Epilepsy Dataset. OpenNeuro. [Dataset]," 2023, doi: doi: doi:10.18112/openneuro.ds004100.v1.1.3.

[19] A. Dosovitskiy et al., "an Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale," ICLR 2021 - 9th Int. Conf. Learn. Represent., 2021.

[20] M. Ott, S. Edunov, D. Grangier, and M. Auli, "Scaling Neural Machine Translation," WMT 2018 - 3rd Conf. Mach. Transl. Proc. Conf., vol. 1, pp. 1–9, 2018, doi: 10.18653/v1/w18-6301.

[21] S. Roy, I. Kiral-Kornek, and S. Harrer, "Chrononet: A deep recurrent neural network for abnormal EEG identification," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11526 LNAI, pp. 47–56, 2019, doi: 10.1007/978-3-030-21642-9_8.

[22] H. RaviPrakash et al., "Deep Learning Provides Exceptional Accuracy to ECoG-Based Functional Language Mapping for Epilepsy Surgery," Front. Neurosci., vol. 14, no. May, pp. 1–14, 2020, doi: 10.3389/fnins.2020.00409.

[23] S. S. Priya and N. J. Nalini, "Eeg Signal Analysis for Epileptic Seizure Using Scaleogram Based Transfer Learning," J. Pharm. Negat. Results, vol. 13, no. 7, pp. 2825–2834, 2022, doi: 10.47750/pnr.2022.13.S07.376.

[24] S. Rukhsar and A. K. Tiwari, "Lightweight convolution transformer for cross-patient seizure detection in multi-channel EEG signals," Comput. Methods Programs Biomed., vol. 242, pp. 1–13, 2023, doi: 10.1016/j.cmpb.2023.107856.

[25] D. Gupta, "Functional iterative approach for universum-based primal twin bounded support vector machine to eeg classification (fuptbsvm)," Multimed. Tools Appl., vol. 83(8), 221, 2024.

[26] Q. Li, T. Zhang, Y. Song, and M. Sun, "Transformer-based spatial-temporal feature learning for P300," 2022 16th ICME Int. Conf. Complex Med. Eng. C. 2022, pp. 310–313, 2022, doi: 10.1109/CME55444.2022.10063297.

[27] Abir Hadriche, Nawel Jmail, Jean-Luc Blanc, Laurent Pezard. "Using centrality measures to extract core pattern of brain dynamics during the resting state" Computer Methods and Programs in Biomedicine, vol. 179. 2019,104985.doi:10.1016/j.cmpb.2019.104985.

Truth Under Pressure: A Deep Learning-Based Lie Detection System for Online Lending Using Voice Stress and Response Latency

Ahmad Ihsan Farhani[1], Alhadi Bustamam[2*], Rinaldi Anwar[3], Dra. Titin Siswantining[4]
Department of Mathematics-Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok, Indonesia [1, 2, 4]
Data Science Center (DSC)-Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok, Indonesia [1, 2]
Global Risk Management (GRM), Jakarta, Indonesia [3]

***Abstract*—The rapid increase in defaults in the online lending industry highlights significant flaws in current debtor verification, which largely relies on static, preparable interviews, leading to high non-performing loans. Existing research is fragmented: while Large Language Models (LLMs) show promise in question generation, their application is confined to non-financial domains like education, and lie detection studies often analyze modalities in isolation. This study addresses this critical gap by proposing the first *integrated* AI-driven system for this context. We solve the problem in two parts: 1) A Llama 3 LLM is fine-tuned to generate dynamic, biodata-tailored questions, preventing the rehearsed answers that plague static interviews. 2) A novel multimodal deep learning model is developed to analyze the response, uniquely fusing vocal acoustic features and response latency—two key deception indicators that prior work has failed to combine. The Llama 3 model produced a low perplexity score (2-3), and the lie detection model achieved 70% testing accuracy with a 70.9% F1-Score. Despite signs of overfitting, this framework provides a novel, intelligent decision-support tool to reduce fraud and manage default risks more effectively.**

***Keywords*—Online lending; lie detection; large language model; deep learning; voice acoustics; response latency**

I. Introduction

In 2024, Indonesia's internet penetration was around 79.50%, which is about 221.5 million people [1]. This high percentage indicates a high level of connectivity and implies a propelling growth in various digital innovations, especially in financial technology (fintech) services. Peer-to-peer (P2P) lending, commonly known as online lending, is one of the most dynamic segments within fintech. This service allows lenders and debtors to transact directly through digital interfaces. Online lending users grew tremendously from 2.7 million to 8.8 million within a year in 2024 [2]. Post-COVID-19 economic realities and a heightened digital consumption mindset have further spurred the adoption of these services [3].

However, the availability of online loans has led to increasing irresponsible borrowing behavior. An increasing number of debtors tended to skip out on easy-to-access loans, which in turn resulted in a rise in the "non-performing loans" (NPLs). In the context of P2P lending, this is measured by the 90-Day Default Rate (TWP90), which refers to loan repayment that has defaulted beyond 90 days [4]. Based on the latest information from the Financial Services Authority of Indonesia, stagnant online loans jumped significantly from IDR 785.94 billion in January 2022 to IDR 1.9 trillion by June 2024 [5]. This is illustrated in Fig. 1.

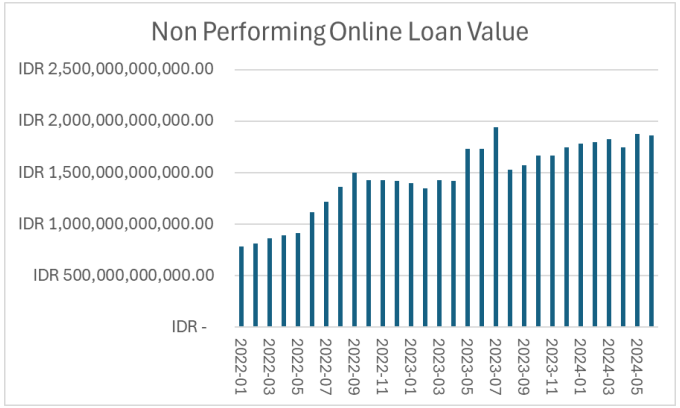


Fig. 1. Accumulated defaulted online loans, 2022–2024.

Moreover, as shown in Fig. 2, the 19 to 34 age group is the largest contributor to loan defaults, followed by individuals aged 35 to 54 [6]. This trend raises serious concerns because the 19 to 34 demographic is the economically productive part of the population that should ideally drive economic growth.

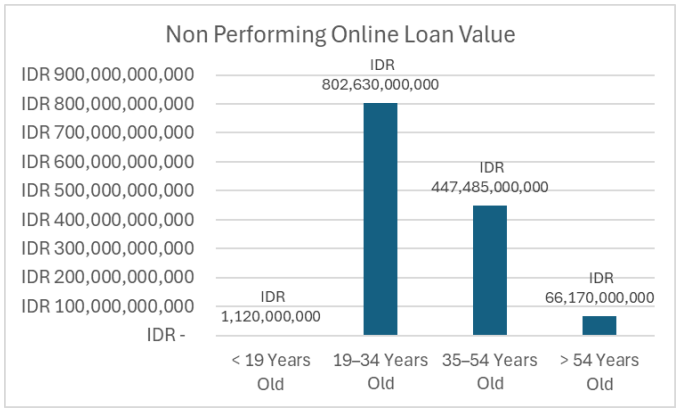


Fig. 2. Default rate by age group.

One of the key factors causing the high rate of loan defaults among the younger productive generations is their

*Corresponding author.

low financial literacy. Several young people use online lending platforms without truly understanding the risks [7]. In 2024–2025, the Financial Services Authority conducted a survey called Indonesia's National Financial Literacy Survey to measure the financial literacy of the Indonesian population. The result of the survey is that in Indonesia, the financial literacy index is only around 65%–66%, and for digital financial literacy, it is at 62% [8],[9].

The survey results from [7]-[9] imply that young people in Indonesia might lack financial understanding, which increases the risk of falling victim to online lending schemes. Several debtors have limited understanding and do not fully understand how online lending works. This causes the debtors to not fully understand the risks and dangers of illegal online lending practices [10]. If this situation continues, the number of non-performing loans will continue to increase. The increase in non-performing loans will weaken personal finances, decrease household consumption, decrease investment, and limit entrepreneurship growth. In the long term, this situation will threaten the national economy.

In addition, verification procedures in online methods of obtaining loans mainly consist of filling in personal data without conducting a direct interview. This opens the opportunity for dishonest debtors to exploit this system. Dishonest debtors can manipulate or change the biodata details they submit, increasing the chances of online loan default. Even in cases where interviews are held, asking static, pre-set questions can be inadequate, as it allows the debtors to prepare rehearsed answers that may not truly represent their financial situation.

To reduce this problem, it is recommended that interview questions be designed to reflect the details of the debtor's biodata so that questions are specific to the debtor and have less predictable answers. However, manual operation on this personalized level would be labor-intensive. In this study, we propose a large language model (LLM)-based automatic interview system and integrate it with a multimodal deep learning architecture.

To provide on-the-spot responses without having to prepare or memorize answers, LLM will use automated question generation (AQG) to generate personalized questions tailored to the debtor biodata. A multimodal model that assesses vocal acoustic characteristics and response latency as possible markers of dishonesty was used to analyze the responses. This system is intended to be implemented before a loan is approved to improve verification accuracy and lower the risk of loan default.

Even though it is still used in the educational domain, previous research has shown the potential of LLMs such as T5, mT5, and Llama in generating questions [11]-[13]. Based on previous research, one of the signs of lying is a change in the acoustic characteristics of speech, such as an increase in pitch and speech disfluency; features such as longer answering process are also one of the deception clues [14]-[16]. All of these signs of deception must be combined to improve model performance as different modalities require a multimodal mechanism. This multimodal model will classify the answer of the debtor as truth or lie.

The major contributions of this study are as follows:

- Proposes a novel integration of vocal acoustic features and response latency to enhance deception detection. These two elements are typically analyzed separately.
- Utilizes a LLM to perform AQG within the financial domain, effectively simulating the online lending interview process.
- Introduces a new verification approach aimed at reducing NPL rates by combining LLM driven questioning and deception detection techniques for a prospective debtor.
- Develops a cost-effective and scalable deception detection framework that supports improved decision-making and operational efficiency in online lending platforms.

This study is organized into five sections. Section I provides the study's background and motivation. Section II explains and shows some related works and previous research relevant to this topic. Section III explains the dataset, methodology, and models used in this study. Section IV presents and discusses the results. Finally, Section V concludes the study and outlines potential directions for future work.

II. Literature Review

This section outlines previous studies related to the current study. This section is divided into two subsections: AQG using LLM and lie detection.

A. AQG Using LLM

AQG using LLM has seen significant growth, particularly in the education domain. As shown in the study [11] and [12], the LLM model with mT5 and T5 architecture has the ability to generate questions that are appropriate to the context of the given input with good human language and good readability. In [11], the authors used the mT5 model trained on the TQuADv2 and XQuAD datasets to generate comprehension questions, achieving BLEU-1 of 47.6 and ROUGE-L of 53.9. Similarly, EduQG, a QG model developed by [12] using T5-small and T5-base, reached BLEU-4 of 15.94 and F1 score of 33.12.

More recently, [13] fine-tuned LLaMA-2 (7B parameters) for question–answer extraction and generation on the SQuAD and PBE datasets, achieving a BLEU-4 of 23.53 and a METEOR of 66.3. Previous studies have shown that LLM can generate dynamic, context-appropriate questions. However, its application is still limited to the educational domain and has not been further explored for adaptation in risk assessments, such as debtor eligibility interviews for online loans.

B. Lie Detection

Several studies have explored lie detection using various modalities such as text, audio, video, facial expression, and multimodal approaches. For instance, [17] implemented a LLM (FLAN-T5) for verbal lie detection using datasets such as Deceptive Opinions, Hippocorpus, and the Intention Dataset, achieving an accuracy of 82.72%. The architecture

created by [18] was named DeepLie, where facial features were extracted using face detection and classified using a deep learning model. The developed DeepLie model achieved a performance of 81.82% on a real-world trial dataset.

Different research approach by [19] extracted high-level features from videos using tools such as OpenFace, part-of-speech n-grams, MFCC, and Covarep and reported a test AUC of 0.730. In [20], the authors combined video, audio (MFCC), and text (GloVe) features and used machine learning classifiers, such as SVM, K-SVM, Decision Tree, and Random Forest, achieving 87.73% accuracy on a trial dataset. Focusing on audio-based detection, [21] developed an acoustic lie detector using MFCC features and models such as SVM and SGD, achieving 62.4% accuracy on the Columbia-SRI-Colorado dataset. In [22], the authors proposed an explainable LSTM-based model that leverages stress analysis in voice signals, achieving an accuracy of 92.4% on a stress voice dataset.

Multimodal deep learning also shows promising results. Research conducted by [23] utilized range of deep learning architecture such as 3D CNNs, text CNNs, and multilayer perceptron and combined them with OpenSMILE for audio features. The research combines video, transcript, and speech to classify a video into a lie or truth class, achieving an impressive accuracy of 97.99% on a trial dataset. Similarly, [24] applied OpenFace and FACS with LSTM for deception detection in videos across datasets such as Silesian, trial, and Bag of Lies, achieving an accuracy of 90.90%.

Another study conducted by [25] used cognitive signals, namely, response latency and error rates, when participants answered unexpected questions as material in detecting lies. Logistic regression, KNN, and SVM were the classifiers used in the study. The highest performance based on the study was an accuracy of 81%, with an AUC of 92%, recall of 80%, precision of 84%, and an F1 score of 80%, highlighting the potential of behavioral timing analysis in lie detection.

C. Research Gap and Proposed Contributions

The literature review reveals significant progress in both AQG and lie detection; however, it also highlights key limitations that this study aims to address.

First, in the field of Automated Question Generation (AQG), the application of powerful LLMs like T5, mT5, and LLaMA is in the educational domain [11]-[13]. There is a distinct lack of research exploring their use in dynamic, high-stakes financial risk assessments, such as debtor interviews.

Second, in Lie Detection, some studies focus solely on a single modality, such as text [17], facial expressions [18], [24], or audio features [21], [22]. Some multimodal systems often combine mostly video, text, and audio either all three of modalities nor combination of 2 modalities [19], [20], [23]. The specific integration of vocal acoustic features and response latency is not explored yet, even though both of the modalities has strong indicators of the cognitive effort associated with deception. Mostly both feature are analyzed separately.

This study directly addresses these gaps. We propose a system that: 1) adapts an LLM for AQG within the financial (online lending) domain and 2) develops a novel, streamlined multimodal model that specifically combine both vocal acoustics and response latency. The following TABLE Isummarizes this research gap and our proposed contributions.

Unlike previous AQG studies that primarily focused on educational or comprehension contexts, our approach introduces domain-specific adaptation and fine-tuning of a large language model for financial risk assessment. The novelty lies in three aspects:

- Prompt engineering strategy, each input prompt integrates structured debtor biodata with contextual financial indicators, enabling the model to generate dynamic and risk-relevant interview questions rather than general comprehension items.
- Domain adaptation, Llama 3 was fine-tuned on a synthetic dataset specifically designed to reflect debtor profiles, borrowing intents, and loan conditions, thus aligning the model's linguistic generation with financial verification contexts.
- Automated reasoning constraint, during decoding, we applied controlled generation parameters (low temperature and limited Top-K sampling) to ensure factual consistency and reduce speculative or irrelevant question output.

Together, these innovations allow the model to perform adaptive, context-aware question generation that simulates human financial interviews a task previously not explored in prior AQG research.

TABLE I. SUMMARY OF RESEARCH GAPS AND PROPOSED CONTRIBUTIONS

Research Area	Identified Limitations / Gaps in Previous Studies	Proposed Solution in This Study
AQG	Application is heavily restricted to the educational domain [11]-[13]. Lacks adaptation for financial risk assessment or debtor interviews.	Develop and fine-tune a LLM model to generate dynamic, personalized interview questions based on debtor biodata specifically for the online lending context.
Lie Detection	Some studies focus on a single modality, such as text [17], facial expressions [18], [24], or audio features [21], [22]. Existing multimodal systems combine video, text, and audio [19], [20], [23]. The specific integration of vocal acoustic features and response latency is not explored. These features are typically analyzed separately.	Propose a novel and efficient multimodal deep learning model that integrates two inputs: vocal acoustic features and response latency.

III. METHODOLOGY

The research methodology follows a multi-stage workflow, as illustrated in Fig. 3, progressing from data acquisition to model integration. The process begins with two

sequential stages: Data Collection and Data Preprocessing. In the collection stage, two distinct datasets are gathered: a synthetic dataset of borrower biodata and interview questions for the AQG task, and the courtroom trial video dataset [26] for lie detection. The subsequent preprocessing stage then cleans, formats, and transforms both the text and audio datasets to make them suitable for training. Following these preparatory steps, the process bifurcates into two distinct, parallel development tracks.

The first track is dedicated to the Automated Question Generation (AQG) component, which involves fine-tuning the Llama 3 model on the prepared biodata dataset. Simultaneously, the second track focuses on building the lie detection system, which entails training the multimodal deep learning model on extracted vocal characteristics and response timing data. Once training is complete, each model is independently assessed in the Model Evaluation stage: the LLM is measured using perplexity and expert review, while the lie detection model is validated using classification accuracy and F1 score. Finally, the two validated models are brought together in the Implementation stage, where they are integrated into a single, cohesive framework for deployment (see Fig. 3).

A. Data Collection

This subsection discusses the data collection process and the datasets used in the study. The dataset used in the research consists of two different data: 1) data for training the LLM model to generate interview questions and 2) data used to train the lie detection model. Each dataset serves a distinct purpose in supporting the development of a comprehensive and intelligent interview system for online lending verification.

1) Interview question dataset: To the author's knowledge, there is no public data containing the debtor's biodata and interview questions that correspond to them, not even the debtor's personal data. This is because such data are sensitive and will not be publicly available. Therefore, this study uses the synthetic data generated using GPT-4o and GPT-4o mini.

The resulting dataset includes 595 borrower profiles and 2,713 generated questions. The data is organized into three key components: Biodata, which contains the prospective debtor's personal and financial information; Analysis, which provides contextual interpretation of the biodata; and Questions, which comprises tailored interview questions generated based on the analysis. A sample of the dataset is shown in 0. This structured format ensures that both question generation tasks are supported by the dataset.

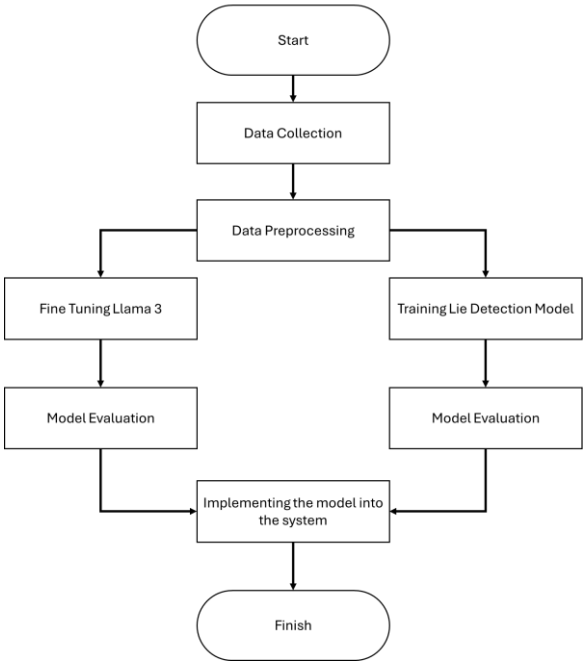


Fig. 3. Methodology of the study. After preprocessing, the fine-tuning of Llama 3 (left) and the training of the lie detection model (right) run in parallel.

TABLE II. SAMPLE BIODATA OF THE QUESTION GENERATION DATASET

Biodata	Analysis	Question
Full Name: Rudi Hartono; **Province**: West Java; **City**: Bekasi; **Educational Background**: High School; **Marital Status**: Single; **Reason for Loan**: Household needs; **Residence**: Rental; **Private Vehicle**: Motorcycle; **Company Name**: PT. Sentosa Abadi; **Occupation**: Private employee Industry; **Sector**: Industry/Factory; **Position**: Supervisor; **Net Monthly Income**: Rp 8,500,000; **Household Monthly Income**: Rp 8,500,000; **Years of Employment**: 4 years; **Emergency Contact**: Parent; **Emergency Contact Name**: Sri Hartono; **Loan Amount**: Rp 15,000,000; **Loan Tenure (Months)**: 18 months; **Monthly Interest Rate**: 6.0%	The reason for borrowing for household needs indicates a shortfall in meeting basic necessities, which may suggest potential financial constraints or inadequate budget management.	1. Do urgent household needs often consume a significant portion of your income? 2. Why do you feel the need to apply for a loan to cover your basic household needs? 3. Do you experience difficulties in managing your monthly finances for daily needs?

2) Lie detection dataset: This study employs the courtroom trial video dataset by [26] for lie detection. The dataset consists of 121 video clips (61 videos are labeled as deceptive class and 60 videos are labeled as truthful class). The videos are labeled based on court evidence and rulings. Audio and these models, which include millions or even billions of parameters, are trained on vast amounts of text data and have demonstrated remarkable capabilities in understanding and generating human-like text. Response latencies were extracted from the videos for analysis. The lie detection dataset compiled from 56 individuals consists of 21 females and 35 males with ages ranging from 16 to 60; durations of the video ranged from 4 seconds to 1 minute 11 seconds. Videos were taken from the recording of real trials; thus, the dataset has varied lighting, camera angles, and poses.

B. Data Preprocessing

This study involves two types of data: text data for the interview generation process and audio recordings for lie detection. Two distinct preprocessing techniques are applied accordingly.

1) Question generation dataset: Tokenization in the Llama 3 model uses a subword-based approach, enabling the Llama 3 model to efficiently handle words not in the model's vocabulary or the presence of affixes. The input text is converted into ID tokens in tokenization. The ID tokens are then converted into mathematical representations through an embedding layer. Each token ID is mapped to a 3072-dimensional vector according to the embedding size used by Llama 3.

Consequently, an input sequence consisting of n tokens is converted into an embedding matrix of size $n \times 3072$. Then, this matrix is used as the input for the transformer layer in the model. The matrix will then be processed in the transformer layer in the model so that the Llama model can understand the input context.

2) Lie detection dataset: To prepare the audio, sound tracks were separated from the video files using the moviepy library. The extracted audio was then adjusted to meet a common format resampled at 16 kHz to keep the data consistent and manageable for processing. Because the recordings varied in length, each clip was truncated and padded into 30 seconds: shorter clips were extended with silent padding, whereas longer clips were trimmed down.

The response latency data were obtained using the pyAudioAnalysis package through a silence removal process. This method segments the audio by removing non-speech regions using a semi-supervised SVM classifier trained to distinguish between high-energy (speech) and low-energy (non-speech) frames. The resulting speech segments were used to calculate the response latency for the lie detection analysis.

3) Data splitting: For robust model assessment and to reduce the likelihood of overfitting, the dataset was separated into two distinct portions: training and testing. The training subset was used to adjust the model parameters. The testing subset supported hyperparameter tuning and helped monitor overfitting during learning.

In this study, stratified random splitting was applied to maintain the balance between the "truth" and "lie" categories across all subsets. This strategy ensured the preservation of class proportions, preventing bias that might otherwise distort model performance. The split was performed using standard library functions, guaranteeing that the process could be replicated consistently. Data allocation followed a widely accepted distribution in machine learning practice: 75% of the data were used for training and 25% for testing.

C. Building Models

1) Llama 3: The primary methodological challenge in the AQG component was adapting the general-purpose Llama 3 model [28] to the specific domain of financial risk assessment. This was not a simple zero-shot prompting task; it required a specific fine-tuning strategy to achieve the desired domain adaptation. We structured our synthetic dataset as 1) into a Biodata, Analysis, Question format. The model was fine-tuned on this structured data to explicitly learn the task of: a) first analyzing a borrower's financial profile and b) then generating contextually relevant, probing questions [27] based on that analysis. This "analysis-then-question" approach is our key technical adaptation, moving the model from a general text generator to a specialized financial risk interrogator.

To execute this efficiently, we utilized Low-Rank Adaptation (LoRA) for fine-tuning, applying it to the model's attention layers. We leveraged the model's native architecture [29] which includes pre-normalization with RMSNorm [30], Rotary Positional Embeddings [31], and the SwiGLU activation function [32] to ensure stable and efficient convergence during our specialized fine-tuning task.

2) SincNet: We selected SincNet as the audio front-end to process raw waveforms directly. This choice was a critical part of our methodology, as traditional hand-crafted features like MFCCs, which are common in audio processing can suppress narrow-band characteristics such as pitch and formants that are key indicators of deception-related stress [33].

SincNet avoids this by using parametrized sinc-based filters in its first convolutional layer [34]. These filters function as learnable band-pass filters defined by cutoff frequencies f_1 and f_2, which are optimized during our training process. The filter is expressed as Eq. (1):

$$g_w[n, f_1, f_2] = (2f_2 \cdot \mathrm{sinc}(2\pi f_2 n - 2f_1 \cdot \mathrm{sinc}(2\pi f_1 n)) \cdot w[n] \quad (1)$$

where, $w[n]$ is a windowing function applied to smooth the filter edges. By implementing this architecture, our model learns the most meaningful frequency features for lie detection directly from the raw audio, rather than relying on pre-engineered features, which is crucial for our multimodal hypothesis.

3) Multimodal: Multimodal is a single algorithm that combines multiple inputs types of input [35]. This study leverages a multimodal deep learning framework to jointly process two types of input: audio data and response latency. Each modality undergoes preprocessing according to its nature. For audio, the signals are resampled, padded, and truncated to achieve a consistent length and dimensionality. Then, SincNet is used to extract the informative features directly from the raw waveform. The silent segments in the audio were identified and removed to capture the response latency, and the resulting features were processed through fully connected layers to produce meaningful numerical representations.

The feature outputs from both modalities are concatenated into a single vector during fusion, forming a shared representation. This combined embedding is then refined through additional fully connected layers to integrate cross-modal information. The final classification layer predicts whether the response reflects truthfulness or deception.

D. Experimental Design

Llama 3 (version 3.2) with 3 billion parameters was employed for the automated question generation task. Fine-tuning was conducted using a learning rate of 2×10^{-4}, over 10 epochs, with a batch size of 3 applied to training and evaluation. During the generation phase, the decoding process was configured with a maximum sequence length of 512 tokens, Top-K sampling set to 50, a temperature of 0.6, and Top-P of 0.95. A multimodal model was trained across 50 epochs with a batch size of 8 for deception detection. The system takes two inputs: raw audio and response latency. Audio signals are processed through SincNet to extract features, whereas response latency is transformed using fully connected layers. Then, the outputs from both pathways were fused for joint representation and final classification.

IV. Experiments and Results

In this section, we explain the results of our research and provide a comprehensive discussion.

A. Evaluation Parameters

1) Question generator evaluation: In this case, the effectiveness of the LLM Llama 3 in generating questions was assessed using perplexity. As described in [36], perplexity quantifies how well a language model or probability distribution predicts a given sample. Lower perplexity values reflect stronger performance, indicating that the model is more confident and precise in forecasting the next token in a sequence. The mathematical formulation of perplexity is expressed as Eq. (2):

$$perplexity = 2^{-\frac{1}{M}\sum_{j=1}^{M}\log_2 \hat{P}(y_j)} \quad (2)$$

2) Lie detection model evaluation: The multimodal deep learning-based lie detection model was assessed using conventional classification metrics derived from the confusion matrix (TABLE III The confusion matrix offers a detailed breakdown of the prediction outcomes by distinguishing between true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [37]. In the context of this binary classification task (Honest vs. Deceptive), the matrix is organized as follows:

Based on [38], the following metrics were calculated from the confusion matrix:

- **Accuracy** reflects the proportion of correctly classified samples relative to the total number of cases and serves as an overall indicator of correctness. The equation is shown as Eq. (3):

$$\text{Accuracy} = \frac{TP + TN}{Total\ Number\ of\ Cases}. \quad (3)$$

- **Precision** denotes the fraction of TPs within all predicted positives, highlighting the effectiveness of the model in minimizing FPs. The equation is shown as Eq. (4):

$$Precision = \frac{TP}{TP\ +\ FP}. \quad (4)$$

- **Recall (or Sensitivity)** measures the share of TPs among all actual positives, illustrating the ability of the model to capture all relevant instances. The equation is shown as Eq. (5):

$$Recall = \frac{TP}{TP\ +\ FN}. \quad (5)$$

- The **F1-score**, defined as the harmonic mean of Precision and Recall, offers a balanced metric that is especially valuable in cases of class imbalance. The equations is shown as Eq. (6):

$$F1 - score = \frac{2 \times p \times r}{p + r} \quad (6)$$

- **Specificity** quantifies the proportion of TNs within all actual negatives, indicating the strength of the model in avoiding false alarms. The equation is shown as Eq. (7):

$$Specificty = \frac{TN}{TN + FP} \quad (7)$$

Together, these metrics provide a comprehensive evaluation framework for assessing the question generation component and the lie detection performance of the proposed system.

TABLE III. Confusion Matrix of Two Classes

		Prediction Label	
		Class 0	**Class 1**
Ground Truth	Class 0	TN	FP
	Class 1	FN	TP

B. Hardware and Software Specification

All experiments in this study were executed using a high-performance computing infrastructure. Table IV shows the hardware and software specifications used for training the LLM and the multimodal deep learning models.

TABLE IV. Hardware and Software Specifications

Hardware	
GPU	RTX 4090
VRAM	24 GB
Ram	128 GB
Storage	1 TB
Software	
Programming Software	VSCode
Programming Language	Python 3.10

C. Performance Analysis

1) Question generator performance: Fig. 4 presents the progression of loss values for training and testing datasets over the course of epochs. During the initial epochs (0–2), training and testing losses sharply declined, indicating that the model quickly captured fundamental patterns and exhibited strong generalization. Between epochs 3 and 6, the training loss continues to decrease. However, after epoch 6, the testing loss levels off and subsequently begins to rise, signaling the

emergence of overfitting. This outcome suggests that the model's large capacity led it to memorize the training data rather than extract broadly applicable features. Approaches such as early stopping and applying regularization methods should be considered to address this issue and preserve generalization.

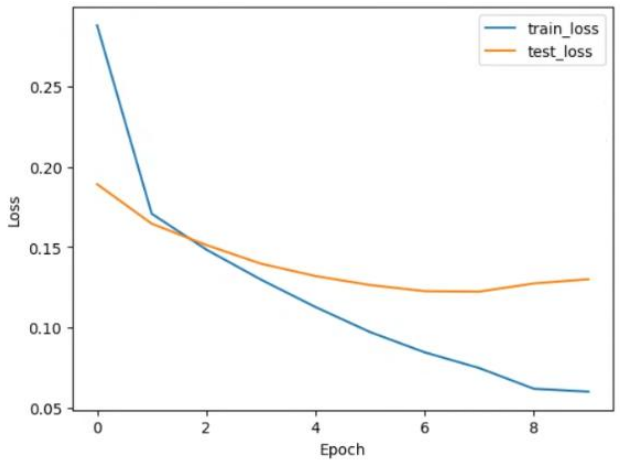


Fig. 4. Changes in the LLM train and test loss with respect to epoch.

The perplexity value for the trained model ranged from 2 to 3, indicating a good predictive capability for subsequent tokens within the language corpus. A low perplexity score in LLMs, such as Llama 3, indicates efficient modeling of the language probability distribution and more accurate predictions. While low perplexity is a positive indicator of performance, the observed overfitting in the test loss highlights the importance of maintaining model generalization for real-world data application.

TABLE Vpresents an example of the output generated by the Llama 3 model. The analysis demonstrates accuracy and professionalism by incorporating details such as estimated monthly installments, their ratio to income, and consideration of rental status as a non-financial risk factor. The generated questions are contextually relevant and exploratory, addressing key aspects such as payment history, emergency fund availability, and preparedness for unforeseen expenses. These results highlight the model's strong potential for producing meaningful analyses and credit-related inquiries, suggesting that it can be effectively integrated into an automated credit scoring system with further refinement.

2) Lie detection performance: This section presents the performance of the lie detection model on the training and testing datasets. The model was trained for 50 epochs with a batch size of 8. Fig. 5 illustrates the evolution of loss values across epochs for both datasets, providing insights into the learning behavior of the model during training and its generalizability to unseen data.

The graph illustrates the progression of the loss values throughout the training of the deep learning model. The horizontal axis represents the number of epochs, whereas the vertical axis shows the loss value. Two curves are displayed: the blue curve corresponds to the training loss and the orange curve represents the testing loss. Both losses are relatively high (approximately 0.7). As training progressed, the training loss dropped steeply, reaching approximately 0.05, reflecting effective learning on the training data. The testing loss also decreased but in a gradual and stable manner, plateauing between epochs 20 and 35 before showing a slight increase toward the end of training. This trend suggests the onset of overfitting after epoch 35, when the model becomes too tailored to the training data, resulting in reduced performance on unseen data.

TABLE V. EXAMPLE OF GENERATED OUTPUT FROM THE TRAINED MODEL

Biodata	Analysis	Questions
Full Name: Rudi Hartono; **Province**: West Java; **City**: Bekasi; **Educational Background**: High School; **Marital Status**: Single; **Reason for Loan**: Household needs; **Residence**: Rental; **Personal Vehicle**: Motorcycle; **Company Name**: PT. Sentosa Abadi; **Occupation**: Private employee; **Industry**: Industry/Factory; **Position**: Supervisor; **Net Monthly Income**: Rp 8,500,000; **Family Monthly Income**: Rp 8,500,000; **Length of Employment**: 4 years; **Emergency Contact**: Parent; **Emergency Contact Name**: Sri Hartono; **Loan Amount**: Rp 15,000,000; **Loan Tenure (Months)**: 18 months; **Loan Interest (Per Month)**: 6.0%	With a loan amount of Rp. 15,000,000 and a tenor of 18 months, the estimated monthly installment is approximately Rp. 1,733,333.33 (with 6% monthly interest). The total repayment will reach Rp. 31,200,000, meaning that around 9.7% of the borrower's monthly income will be allocated for loan repayment.	1. Have you ever been late in paying installments or loans before? 2. What is your plan to ensure that this installment will not interfere with your routine expenses? 3. Do you currently have any other ongoing loans or credits? If yes, what is the total monthly installment you need to pay?

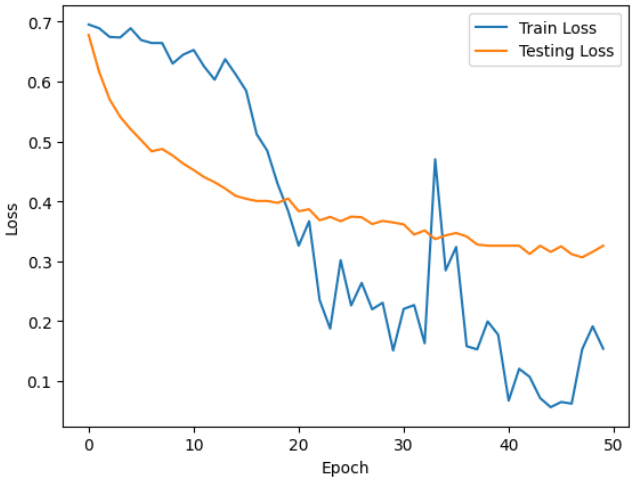


Fig. 5. Change in the lie detection model train and test loss with respect to epoch.

Fig. 6 depicts the evolution of the model's accuracy during the 50-epoch training process, with the horizontal axis representing epochs and the vertical axis showing the accuracy

in percentage. There are two curves: the blue curve represents the training accuracy, whereas the orange curve shows the testing accuracy. At the beginning of training, the training accuracy started at a relatively low and fluctuating value, but generally increased sharply to over 90%. Despite this increase, the training accuracy curve shows considerable fluctuations at several points, likely owing to unstable training data or varying batches.

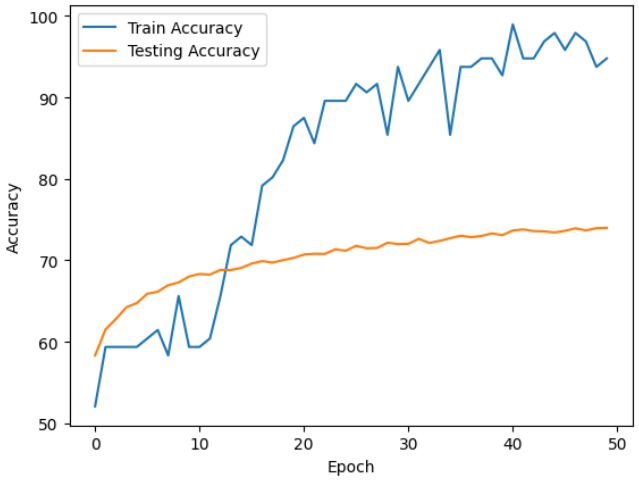


Fig. 6. Change in the lie detection model train and test accuracy with respect to epoch.

Conversely, the testing accuracy curve appeared smoother and more consistent, showing a gradual increase from approximately 58% to nearly 70%, before stabilizing toward the end of training. The substantial gap between the training and testing accuracy curves highlights the presence of significant overfitting. This discrepancy indicates that although the model achieved outstanding performance on the training data, it struggled to effectively generalize to unseen data.

The developed lie detection model achieved 100% accuracy on the training set (TABLE VIand TABLE VIIThis outcome shows that the model successfully captured the underlying patterns of the training data. However, when evaluated on the test set, which contained an evenly balanced distribution of the "honest" (class 0) and "deceptive" (class 1) classes, its accuracy dropped to 70% (TABLE VI. This decline in performance further confirms the occurrence of overfitting, indicating that while the model excelled in memorizing the training data, it exhibited limited generalization capability when applied to new, unseen samples.

TABLE VI. CONFUSION MATRIX OF THE LIE DETECTION MODEL ON TRAINING DATA

	Class	Prediction	
		Honest	Lie
Ground Truth	Honest	45	0
	Lie	0	46

TABLE VII. PERFORMANCE OF THE LIE DETECTION MODEL ON TRAINING DATA

	Accuracy	Precision	Recall	F1 Score	Specificity
Train	100%	100%	100%	100%	100%

Based on the confusion matrix in TABLE VIIIand TABLE IXfor the testing data, 10 honest samples were correctly classified and 5 honest samples were misclassified as deceptive (FP). Of the 15 deceptive samples, 11 were correctly classified, whereas 4 others were misidentified as honest (FN). The precision and recall values for the deceptive class were 68.8% and 73.3%, respectively. This relatively small difference indicates that the model has a balanced performance in detecting deception, both in terms of its ability to correctly identify deceptive cases (recall) and the accuracy of predicting deception (precision). However, the slightly lower precision compared with recall suggests that there are still some incorrect deceptive predictions (FPs), although the overall performance of the model can be considered good in detecting deception. Specificity, the ability of the model to correctly detect honesty, was 66.67%.

TABLE VIII. CONFUSION MATRIX OF THE LIE DETECTION MODEL ON TESTING DATA

	Class	Prediction	
		Honest	Lie
Ground Truth	Honest	10	5
	Lie	5	11

TABLE IX. PERFORMANCE OF THE LIE DETECTION MODEL ON TESTING DATA

	Accuracy	Precision	Recall	F1 Score	Specificity
Train	70%	68.8%	73.3%	70.9%	66.67%

In a real-world context, misclassifying an honest person as deceptive can have a negative impact on reputation and trust. Conversely, if a deceptive person goes undetected as honest, other forms of loss can occur. Therefore, a better balance between the model's sensitivity (recall) and specificity is required to improve its overall performance.

D. Discussion

The experimental results of this study demonstrate the potential and current limitations of employing artificial intelligence to improve verification processes in online lending, particularly through automated lie detection. The proposed system, which integrates a LLM for personalized question generation and a multimodal deep learning model for deception detection, represents a novel approach to reducing the risk of loan default.

The LLM-based question generator, built upon Llama 3, successfully generated contextually relevant and individualized questions by analyzing borrower biodata. The observed perplexity scores between 2 and 3 indicate that the model learned meaningful representations of the input data and could effectively predict token sequences. Nonetheless, the training dynamics revealed signs of overfitting,

particularly beyond epoch 6, where the testing loss began to increase despite continuous improvement in the training loss. This phenomenon suggests that the model began to memorize training samples rather than learning generalized patterns. Future implementations should consider employing regularization techniques such as dropout as well as strategies such as early stopping and k-fold cross-validation to mitigate this issue.

The lie detection model demonstrated strong learning ability during training, achieving 100% accuracy. However, the performance on the testing dataset declined to 70% accuracy, with additional metrics including a precision of 68.8%, recall of 73.3%, F1 score of 70.9%, and specificity of 66.67%. These figures reflect a moderate level of generalization and indicate that the model was more proficient in detecting deceptive behavior (as shown by its relatively higher recall) than in avoiding FPs. This trade-off between sensitivity and specificity is particularly important in financial contexts: while failing to detect deception may result in financial loss, false identification of honest applicants can damage user trust and accessibility.

The discrepancy between the training and testing performance, along with the presence of FPs and FNs, indicates the need for further refinement of the model. The domain gap between the courtroom trial audio data used for training and the real-world voice patterns of loan applicants may be contributing factors. Therefore, the collection and use of domain-specific audio data, sourced directly from online lending environments, is recommended to improve model adaptability and contextual accuracy.

Moreover, while the technical feasibility of the proposed system is evident, its practical deployment raises several ethical concerns. The automated classification of debtor honesty based on vocal features and response timing may introduce biases, especially for individuals with speech disorders, anxiety, or cultural-linguistic variations. These risks necessitate the implementation of fairness-aware machine learning techniques and the potential integration of a human-in-the-loop review mechanism.

In summary, the results confirm that the integration of adaptive LLM-based questioning with voice and response latency analysis holds substantial promise for improving the objectivity and robustness of digital credit verification systems. However, to transition this system from proof-of-concept to deployment, further work is required in terms of data expansion, overfitting mitigation, domain adaptation, and ethical governance. This study's methodology provides a solid foundation for the future development of scalable, intelligent decision-support tools for financial technology platforms.

V. Conclusion

This study presents an AI-driven lie detection system tailored for online lending applications, integrating a LLM with a multimodal deep learning framework. The proposed system aims to improve the borrower verification process by generating personalized, context-aware interview questions using Llama 3 and subsequently evaluating applicant honesty through vocal acoustic features and response latency.

The experimental results show that the system performs well during the training phases, achieving 100% accuracy in lie detection on the training data and producing highly relevant, individualized questions. However, generalization remains a challenge, as evidenced by the drop in testing accuracy to 70% and indications of overfitting in the question generation and deception detection components. Despite this, the model achieved a balanced performance on unseen data, with an F1 score of 70.9%, precision of 68.8%, recall of 73.3%, and specificity of 66.67%.

This integrated framework represents a significant advancement over existing methods. Unlike traditional, static questionnaires, the LLM-driven interview prevents rehearsed answers; unlike rule-based scoring, it adds behavioral analysis; and unlike unimodal models, its fusion of vocal acoustics and latency provides a more robust deception signal. This scalable framework thus shows strong potential as a decision-support tool. Its automation enables thousands of simultaneous interviews, reducing labor costs and financial losses from NPLs, with broader implications for recruitment, insurance, and fraud investigations.

Future research should focus on mitigating overfitting through hyperparameter tuning, applying cross-validation techniques, expanding and diversifying the dataset, and incorporating advanced feature extraction methods. Furthermore, ethical considerations, including bias mitigation and human oversight, must be integrated into the deployment pipeline to ensure fairness and transparency in real-world applications.

Acknowledgment

This research was supported by the Program Penelitian Tahun Anggaran 2025, under the Directorate of Research and Community Service, Directorate General of Research and Development (DIKTI), Ministry of Higher Education, Science, and Technology of Indonesia, through Research Grant No. PKS-507/UN2.RST/HKP.05.00/2025. The authors also acknowledge the additional support provided by PT. Global Risk Management (GRM), PT. Seleris Meditekno Internasional, Data Science Center (DSC), and the Laboratory of Bioinformatics and Advanced Computing, Department of Mathematics, Universitas Indonesia for the successful completion of this research.

References

[1] Asosiasi Penyedia Jasa Internet Indonesia, "Hasil survei internet APJII.", *Online*, Nov. 2024.

[2] D. Wati and T. Syahfitri, "Dampak pinjaman online bagi masyarakat," *Community Dev. J.*, vol. 2, no. 3, pp. 1181–1186, Nov. 2022.

[3] J. Z. Y. Arvante, "Dampak permasalahan pinjaman online dan perlindungan hukum bagi konsumen pinjaman online," *Ikatan Penulis Mahasiswa Hukum Indones. Law J.*, vol. 2, no. 1, pp. 73–87, Feb. 2022.

[4] R. Kartika and M. Umam, "Tingkat wanprestasi 90 peer to peer lending selama pandemi COVID-19 di Indonesia," *Akuntabilitas J. Ilm. Ilmu-Ilmu Ekon.*, vol. 14, no. 1, pp. 31–40, Nov. 2021.

[5] A. Ahdiat, "Kredit macet pinjol tembus Rp1,9 triliun akhir semester I 2024," *Katadata Databoks*. Nov. 2024.

[6] C. M. Annur, "Nilai kredit macet pinjol berdasarkan kelompok usia penerima pinjaman," *Katadata Databoks*. Aug. 2023.

[7] P. F. Andini, I. Zidane, H. D. Wahyuni, M. Y. I. Rabbani, and L. N. Yuliati, "Perlindungan konsumen pinjaman online yang mengalami

keterlambatan pembayaran cicilan," *Policy Brief Pertan. Kelautan Biosains Trop.*, vol. 4, no. 4, pp. 381–386, Feb. 2022.

[8] U. S. Liman, "AFPI memanfaatkan kecerdasan buatan cegah penipuan pinjaman daring," *Antara News*, May 6, 2025.

[9] M. H. Simanjuntak, "OJK: Literasi keuangan digital cegah masyarakat dari pinjol ilegal," *Antara News*, Nov. 4, 2024.

[10] M. Suhayati, "FENOMENA KREDIT MACET PINJAMAN ONLINE," *Isu Sepekan Bidang Ekkuinbang, Komisi XI,* Jul. 2023.

[11] F. C. Akyon, D. Cavusoglu, C. Cengiz, S. O. Altinuc, and A. Temizel, "Automated question generation and question answering from Turkish texts using text-to-text transformers," *arXiv preprint,* Nov. 2021.

[12] S. Bulathwela, H. Muse, and E. Yilmaz, "Scalable educational question generation with pre-trained language models," *arXiv preprint,* Apr. 2022.

[13] G. H. Alferez, "Comprehensive question and answer generation with LLaMA 2," *M.S. thesis*, Southern Adventist Univ., Tennessee, 2024.

[14] P. Ekman, M. O'Sullivan, W. V. Friesen, and K. R. Scherer, "Face, voice, and body in detecting deceit," *J. Nonverbal Behav.*, vol. 15, no. 2, pp. 125–135, Jun. 1991.

[15] C. Kirchhuebel, "The acoustic and temporal characteristics of deceptive speech," *Ph.D. dissertation*, Univ. of York, York, U.K., Oct. 2013.

[16] B. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," *Psychol. Bull.*, vol. 129, no. 1, pp. 74–118, Jan. 2003.

[17] R. Loconte, R. Russo, P. Capuozzo, P. Pietrini, and G. Sartori, "Verbal lie detection using large language models," *Sci. Rep.*, vol. 13, no. 1, p. 22849, Dec. 2023.

[18] J. Feng, "DeepLie: Detect lies with facial expression (computer vision)," 2021.

[19] R. Rill-García, H. J. Escalante, L. Villaseñor-Pineda, and V. Reyes-Meza, "High-level features for multimodal deception detection in videos," *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, pp. 1565–1573, Jun. 2019.

[20] Z. Wu, B. Singh, L. Davis, and V. Subrahmanian, "Deception detection in videos", *AAAI*, vol. 32, no. 1, p. 11502, Apr. 2018.

[21] H. Rohde and P. A. Finkelstein, "An acoustic automated lie detector," *Semant. Scholar Corpus*, Jan. 2019.

[22] F. M. Talaat, "Explainable enhanced recurrent neural network for lie detection using voice stress analysis," *Multimed. Tools Appl.*, vol. 83, no. 11, pp. 32277–32299, Nov. 2024.

[23] G. Krishnamurthy, N. Majumder, S. Poria, and E. Cambria, "A deep learning approach for multimodal deception detection," *arXiv preprint*, Mar. 2018.

[24] H. U. D. Ahmed Khan, U. I. Bajwa, N. I. Ratyal, F. Zhang, and M. W. Anwar, "Deception detection in videos using the facial action coding system," *Multimed. Tools Appl.*, vol. 84, no. 9, pp. 6429–6443, Apr. 2024.

[25] G. Melis, M. Ursino, C. Scarpazza, A. Zangrossi, and G. Sartori, "Detecting lies in investigative interviews through the analysis of response latencies and error rates to unexpected questions," *Sci. Rep.*, vol. 14, no. 1, p. 12268, May. 2024.

[26] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. pp. 59–66, Nov. 2015,

[27] M. Ridho, A. Bustamam, and R. Adnan, "Reconstruction of the Phi-2 Method for Question-Answering Related to Diabetes Disease Using the MedAlpaca Dataset," Jambura Journal of Biomathematics (JJBM), vol. 6, no. 3, pp. 183–187, 2025

[28] A. Dubey *et al.*, "The LLaMA 3 herd of models,", *arXiv preprint*. Nov. 2024.

[29] M. Tamang, "Build your own Llama 3 architecture from scratch using PyTorch," *Towards AI*. Sep. 2024.

[30] B. Zhang and R. Sennrich, "Root mean square layer normalization," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst. (NeurIPS).* 2019.

[31] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "RoFormer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, Feb. 2024.

[32] N. Shazeer, "GLU Variants Improve Transformer," *arXiv preprint,* Feb 2020.

[33] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *Proc. Interspeech*, Aug. 2017.

[34] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with SincNet," *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028, Jul. 2018.

[35] N. Hilmizen, A. Bustamam and D. Sarwinda, "The Multimodal Deep Learning for Diagnosing COVID-19 Pneumonia from Chest CT-Scan and X-Ray Images," *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, Yogyakarta, Indonesia, pp. 26-31, 2020.

[36] T. B. Brown *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, May. 2020.

[37] J. Patterson and A. Gibson, *Deep learning: A practitioner's approach*, 1st ed. Sebastopol, CA, USA: O'Reilly Media, Inc., 2017.

[38] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *Int. J. Data Min. Knowl. Manag. Process*, vol. 5, pp. 1–11, Mar. 2015.

Classification of Mangrove Ecosystem Health Using Sentinel-2 Images with Genetic Algorithm Optimization in Machine Learning Algorithms

Putri Yuli Utami[1]*, Murni Ramadhani[2], Rudi Alfian[3], Barry Ceasar Octariadi[4], Dimas Kurniawan[5]
Information System-Faculty of Engineering and Computer Science, Universitas Muhammadiyah Pontianak, 78124 Indonesia[1, 2, 5]
Aquaculture-Faculty of Fisheries and Marine Science, Universitas Muhammadiyah Pontianak, 78124 Indonesia[3]
Informatics-Faculty of Engineering and Computer Science, Universitas Muhammadiyah Pontianak, 78124 Indonesia[4]

***Abstract*—Mangrove ecosystems play an important role in maintaining coastal ecological balance, including as carbon sinks and natural protection from abrasion, but mangrove areas in Mempawah Regency have experienced significant degradation due to anthropogenic pressures. Therefore, this study aims to classify the health condition of mangroves using multi-temporal Sentinel-2 imagery with a hybrid machine learning (ML) approach and Genetic Algorithm (GA) optimization. We implemented GA optimization comparatively on four main ML models—Multilayer Perceptron (MLP), Decision Tree (DT), XGBoost, and Naïve Bayes (NB)—to adjust hyperparameters to improve accuracy and reduce overfitting. The results prove that GA optimization effectively improves classification performance, with the MLP-GA model providing the highest accuracy with an increase of up to 3.8% compared to the non-optimized baseline model, achieving a best performance value of ROC AUC 0.9730 and reducing computation time by up to 60%. These findings indicate that the GA-MLP framework is highly reliable and efficient, providing a precise tool for strategic decision-making in the management of healthy mangrove ecosystems.**

***Keywords*—Classification; genetic algorithm; machine learning; mangrove ecosystem; Sentinel-2**

I. Introduction

Mangrove forests cover only about 0.7% of the world's total tropical forests [1], but has the ability to store three to four times more carbon per unit area than other tropical forests, and in areas with carbonate and peat environments, their storage capacity can be 25 to 50% higher [2]. Globally, this ecosystem is spread across more than 118 countries with an area of approximately 152,000 km² [3]. Meanwhile, in Indonesia, mangroves are found throughout the archipelago. The largest distribution is in Irian Jaya, covering an area of 1,350,600 ha (38%), followed by Kalimantan with 978,200 ha (28%), and Sumatra with 673,300 ha (19%) [4][5]. Meanwhile, in other regions, mangroves generally grow optimally on coasts with large river estuaries and protected beaches.

West Kalimantan is a province with a coastline stretching 1,940 km, most of which is dominated by mangrove ecosystems [6][7]. The area of mangrove forests in this province is estimated to reach 119,327 ha, with a composition of around 75% of the mangrove species found in Indonesia. On Padang Tikar Island, Batu Ampar District, Kubu Raya Regency, the mangrove ecosystem covers an area of 58,953 ha. Meanwhile, Mempawah Regency, which is also located on the coast of West Kalimantan, has a fairly high diversity of mangroves [8]. However, in the last 25 years, it has experienced degradation covering an area of 250.11 ha, so that in 2014, only about 739.31 ha remained [9]. This condition indicates that the mangrove ecosystem in the study area not only has high ecological value but also faces serious threats, making remote sensing-based mangrove health monitoring urgently needed.

Monitoring the health of mangrove ecosystems is crucial to support conservation and rehabilitation strategies. The use of remote sensing technology, particularly Sentinel-2 satellite imagery, offers great opportunities because it has sufficient spatial, temporal, and spectral resolution to map land cover conditions and detect vegetation health dynamics [10][11][12]. Several vegetation and water indices, such as NDVI (Normalized Difference Vegetation Index) [13][14], SAVI (Soil-Adjusted Vegetation Index) [15], NDWI (Normalized Difference Water Index) [16], and MNDWI (Modified Normalized Difference Water Index) [17], have proven effective in evaluating vegetation density, humidity, and coastal ecosystem conditions. However, challenges still arise in terms of classification accuracy and consistency due to the complexity of mangrove spectral characteristics, which often overlap with other vegetation.

Previous studies have shown various approaches to mangrove mapping and monitoring. The XGBR-GA hybrid approach, which combines Extreme Gradient Boosting Regression and Genetic Algorithm for feature selection with multi-source data (Sentinel-2, Sentinel-1, ALOS-2 PALSAR-2) and field data (105 plots), successfully estimated Above-Ground Biomass (AGB) with good accuracy (R^2 = 0.683; RMSE = 25.08 $Mg \cdot ha^{-1}$) [18]. Furthermore, the EIAGA-S (Elite Individual Adaptive Genetic Algorithm-Semantic Inference) method, combined with the new MEVI index, was able to perform semantic segmentation of mangroves without ground truth with high accuracy (mIoU = 0.92; F1 = 0.923) and multi-class classification. On the other hand, NDVI-based research in Tugu District, Semarang, shows that the mangrove area covers 113.93 ha with a health condition dominated by the "Good" category (57.1 ha), although the effectiveness of mangroves in mitigating coastal abrasion (970 m in 2017–2021) is still limited [19]. The three studies introduced the Elite Individual Adaptive Genetic Algorithm-Semantic Inference (EIAGA-S) method for semantic segmentation of mangroves

* Corresponding author.

without requiring ground truth data. This method combines an adaptive genetic algorithm with an elite evolution strategy and develops the Mangrove Enhanced Vegetation Index (MEVI) to distinguish mangroves from other vegetation. Evaluation results on multi-source data show high performance with mIoU = 0.92 and F1-score = 0.923, surpassing traditional models such as K-means and SVM [20].

The development of machine learning methods has opened up new opportunities to improve the accuracy of land cover classification. Algorithms such as Multilayer Perceptron (MLP) excel at capturing nonlinear patterns [21], Extreme Gradient Boosting (XGBoost) is powerful in processing complex data [21], Easy-to-interpret Decision Tree [22], and Naïve Bayes, which is efficient for text data, has been widely used in satellite image analysis, but its performance is greatly influenced by the selection of complex hyperparameters [23]. To address these challenges, this study integrates Genetic Algorithm (GA) as an optimization method to improve accuracy while minimizing the risk of overfitting, so that the resulting model is more reliable in supporting coastal ecosystem monitoring [24][25].

Data processing was carried out using Sentinel-2 imagery from the COPERNICUS/S2 HARMONIZED collection accessed through Google Earth Engine (GEE), with the study area focused on the coast of Mempawah Regency, West Kalimantan, which is bounded by a polygon-shaped AOI. To monitor the temporal dynamics of mangrove ecosystem health, the images were analyzed over three time periods (2019–2021, 2021–2023, and 2023–2025), then several spectral indices such as NDVI, SAVI, NDWI, and MNDWI were calculated to assess the condition of mangrove vegetation and detect changes in land cover from year to year.

This study is expected to make a significant methodological and practical contribution. The methodological contribution lies in the comparative evaluation of Genetic Algorithm (GA) optimization in four machine learning models for mangrove health classification, which has rarely been explored in the context of multi-temporal Sentinel-2 data. Quantitatively, this study proves that GA can improve the accuracy of the best model (MLP) by up to 3.8% compared to the non-optimized base model, achieving the highest ROC AUC performance value of 0.9730, while reducing hyperparameter computation time by up to 60%. These precise data-based results are crucial for supporting conservation and rehabilitation strategies. Furthermore, this study aligns with the UN's 2030 Agenda, particularly Sustainable Development Goal (SDG) 14.2 [26], which emphasizes the protection and sustainable use of marine and coastal ecosystems, making healthy mangrove management key to supporting sustainable development. Overall, this study provides recommendations that can be directly applied in mangrove reforestation efforts, supporting sustainable development (SDGs 2030) through more targeted mangrove planting strategies to enhance ecological resilience in coastal areas [27].

The remainder of this study is organized as follows: Section II presents a review of various related works, highlighting the limitations of previous studies and explaining how this research addresses these gaps. Section III details the methods used, ranging from Sentinel-2 image processing, multi-index data preparation, and the implementation of Genetic Algorithm (GA) optimization on four machine learning models. Section IV presents the experimental results, and an in-depth discussion of the optimized model performance comparison is presented in Section V. Finally, Section VI summarizes the main conclusions of this study and provides directions for future work.

II. Related Work

Remote sensing-based mangrove ecosystem monitoring research has been conducted extensively using various approaches. One of the most widely used methods is vegetation indices, such as NDVI, SAVI, NDWI, and MNDWI, which have proven effective in assessing vegetation density, moisture conditions, and distinguishing between vegetative and non-vegetative areas. For example, research in the Semarang area successfully identified the extent of mangroves and classified their health conditions with a reasonable degree of accuracy [28]. However, the study was limited to a single observation period, so it was not able to describe changes in mangrove conditions over time.

With advances in data analysis technology, machine learning algorithms are increasingly being applied to improve the accuracy of satellite image classification. Various studies have demonstrated the success of algorithms such as Multilayer Perceptron (MLP), Extreme Gradient Boosting (XGBoost), Decision Tree, and Naïve Bayes in land cover mapping and vegetation identification. MLP is known to be effective in capturing nonlinear patterns between spectral variables [29], whereas XGBoost excels at handling complex data and is able to overcome overfitting issues through its efficient gradient boosting mechanism [30][31]. Decision Trees are often used because they are easy to interpret and capable of displaying transparent classification rules [32]. Meanwhile, Naïve Bayes has an advantage in computational efficiency for high-dimensional data thanks to its simple yet effective assumption of feature independence [33]. However, the performance of these algorithms is highly dependent on the selection of appropriate hyperparameters. Manual parameter tuning is often inefficient and can reduce the accuracy and stability of the model, especially with complex remote sensing data [34][35].

To overcome problems in hyperparameter tuning, various studies have begun to apply metaheuristic-based optimization algorithms such as the Genetic Algorithm (GA). One of these studies was conducted by Tien Dat Pham et al. (2022), who developed the XGBR-GA model for estimating above-ground biomass (AGB) of mangroves. This model utilizes data from Sentinel-1, Sentinel-2, ALOS-2 PALSAR-2, as well as field data from 105 plots in the Red River Delta, Vietnam. The results of the study show the best performance with an R^2 value of 0.683 and an RMSE of 25.08 $Mg \cdot ha^{-1}$, making the XGBR-GA approach effective for monitoring mangrove ecosystems in tropical regions [36]. In addition, research by Xinhong Li et al. (2023) developed an explainable machine learning method for Fractional Vegetation Cover (FVC) inversion in the alpine grasslands of the Qinghai-Tibet Plateau. A combination of Genetic Algorithm (GA), XGBoost, and Optuna (GA-OP) was used for feature selection and hyperparameter tuning, resulting in a Stacking model with the best performance (R^2 = 0.867;

RMSE = 0.12). The SHAP method and NDVI-CV analysis improved the interpretability and reliability of the results, making this approach effective for estimating other ecological parameters [37]. These results demonstrate the effectiveness of GA in improving the accuracy, stability, and generalization ability of machine learning models for large-scale remote sensing data.

Although a number of previous studies have successfully improved the accuracy of mangrove classification through the application of machine learning algorithms and Genetic Algorithm (GA)-based optimization, most of these studies still have several important limitations. First, most previous studies have focused on a single observation period, thus failing to consider the temporal dynamics of mangrove ecosystems, which are greatly influenced by seasonal factors and anthropogenic activities. Second, the study areas tend to be limited to small and homogeneous areas, so the results cannot describe the complex spatial variations in larger tropical areas. Third, some studies only rely on one or two vegetation indices (such as NDVI or SAVI) without considering water indices such as NDWI and MNDWI, even though these two indices are important for assessing the interaction between mangrove vegetation and the surrounding aquatic environment. Fourth, the GA approach used is generally only applied to one machine learning model, so there has been no discussion of the effectiveness of multi-model or cross-algorithm optimization.

This study attempts to overcome these limitations through a more comprehensive and adaptive approach to classifying mangrove ecosystem health. First, a multi-temporal analysis was conducted for the period 2019 to 2025 to describe the dynamics of changes in mangrove health over time. Second, the study area was focused on the coast of Mempawah Regency, West Kalimantan—an area with high vegetation diversity and real environmental pressures—so that it could represent more diverse ecological conditions. Third, this study combined four main indices (NDVI, SAVI, NDWI, and MNDWI) to provide a more comprehensive spatial picture of vegetation and water conditions. Fourth, the novelty of this study lies in the application of GA across four machine learning algorithms (MLP, XGBoost, Decision Tree, and Naïve Bayes), which are simultaneously optimized to improve the accuracy, stability, and generalization ability of the model. Thus, this study not only expands the spatial and temporal coverage but also introduces methodological innovations in the application of GA for more effective and scientific classification of mangrove ecosystems.

III. Methodology

This research method consists of nine main stages for land cover classification and mangrove health assessment based on Sentinel-2 imagery. The stages include determining the study area, acquiring Sentinel-2 Level-2A imagery, pre-processing (cloud masking and extraction of six spectral bands), and creating a labeled dataset through pixel value extraction from sample points. Five machine learning algorithms (MLP, Naive Bayes, Decision Tree, and XGBoost) were applied with parameter optimization using Genetic Algorithm. Model evaluation was performed using accuracy, precision, recall, F1-score, and Cohen's Kappa, then the best model was used to produce a mangrove cover classification map and interpretation of its health condition. The overall research process flow, from image acquisition to result interpretation, can be seen visually in Fig. 1.

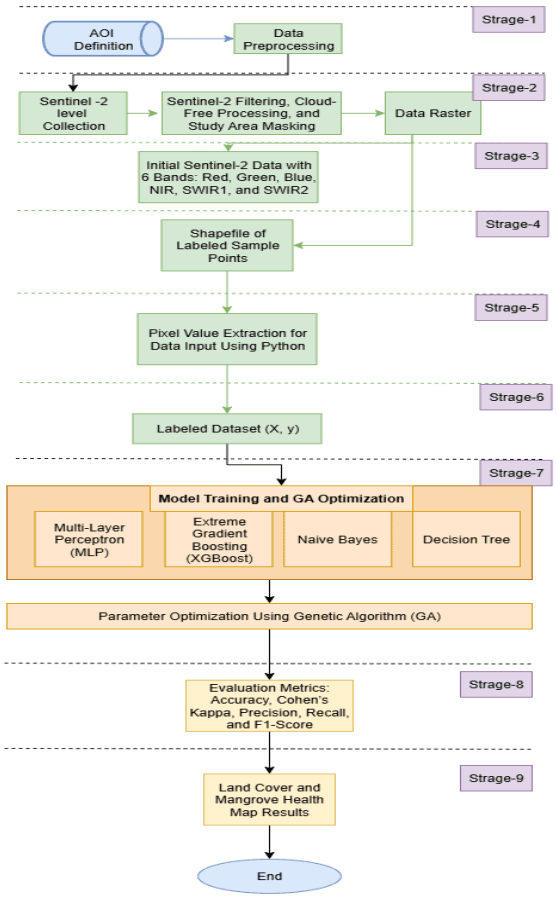


Fig. 1. Research process flow.

A. Stage-1: Determination of AOI and Data Preparation

The initial stage of the study began with the determination of the Area of Interest (AOI) located in the coastal area of Mempawah, West Kalimantan. This AOI was defined using polygon vector data with a two-dimensional geographic coordinate system (Geographic 2D CRS: EPSG:4326 – WGS 84), covering an area with longitude coordinates ranging from 108.829975° to 109.043271° and latitude coordinates ranging from 0.295701° to 0.565236°. The projection system used is WGS 84 (World Geodetic System 1984) with an ellipsoidal datum and Greenwich prime meridian. This shapefile data is then used to determine sample points in the class labeling process. The base image used is Sentinel-2 Level-2A, which was collected through the Google Earth Engine (GEE) platform and will be used in the spectral value extraction stage.

B. Stage-2: Sentinel-2 Image Processing

The Sentinel-2 images collected were then processed through several stages, including masking the study area, cloud filtering, and selecting the study area to produce a raster ready for analysis. This study uses Sentinel-2 images from the COPERNICUS/S2 HARMONIZED collection available on the Google Earth Engine (GEE) platform

(https://code.earthengine.google.com/), with a cloud cover filter of less than 20%. The study area is focused on the coastal region of Mempawah Regency, West Kalimantan, and is bounded by a polygonal AOI.

The image selection process was carried out for three time periods, namely: 2019 to 2021, 2021 to 2023, and 2023 to 2025, in order to monitor the dynamics of mangrove cover changes over time. The selection of these three time periods aims to monitor the temporal dynamics of changes in mangrove ecosystem health, thereby providing an overview of degradation and recovery trends from year to year.

Each image that meets the criteria is then calculated for its spectral index value, such as NDVI, SAVI, MNDWI, and NDWI, to identify the health level of mangrove vegetation. The processing results are stored in TIFF format. AOI in the form of a shapefile with the ID asset projects/mengrove-riset-24/assets/AOI is accessed through the GEE API and converted into FeatureCollection, with geometric information obtained through the getInfo() method.

C. Stage-3: Spectral Band Extraction

The initial Sentinel-2 image was extracted using six important bands, namely Red, Green, Blue, NIR, SWIR1, and SWIR2. These bands were selected because they are relevant for detecting vegetation and plant health, especially in mangrove ecosystems.

TABLE I VEGETATION AND WATER INDICES FROM SENTINEL-2 IMAGES

Vegetation Indices	Formula	*Sources*
Blue	B2	-
Green	B3	-
Red	B4	-
NIR	B8	-
SWIR-1	B11	-
SWIR-2	B2	-
NDVI (*Normalized Difference Vegetation Index*)	(NIR - Red) / (NIR + Red)	[38]
SAVI (*Soil-Adjusted Vegetation Index*)	((NIR - Red) / (NIR + Red + L)) * (1 + L)	[39]
MNDWI (*Modified Differential Water Index*)	(Green - SWIR1) / (Green + SWIR1)	[40]
NDWI (*Normalized Difference Water Index*)	(Green - NIR) / (Green + NIR)	[41]

As presented in Table I, various vegetation and water indices were calculated from a combination of Sentinel-2 spectral bands to assess vegetation cover and water conditions. Four main indices were used in this study, namely NDVI, SAVI, NDWI, and MNDWI. NDVI is widely used to measure vegetation density and greenness, making it a relevant indicator of mangrove health. SAVI was developed to minimize the influence of soil background, making it more accurate in areas with sparse or uneven vegetation cover, including mangrove ecosystems in coastal areas. Meanwhile, NDWI and MNDWI are used to detect vegetation moisture and distinguish water areas from dry or built-up land—important parameters in mangrove ecosystems that are highly influenced by tidal and water conditions. Thus, the combination of these indices provides a more comprehensive spatial picture of mangrove health status.

D. Stage-4: Sample Labeling

Sample labeling was carried out using points from a shapefile that had been prepared beforehand. These points represent land cover classes and mangrove health conditions that were determined manually through visual interpretation and field verification. In this study, 1,250 sample points were used, divided evenly into five land use classes, each with 250 points, namely: 1) Water Body, 2) Non-Mangrove Vegetation, 3) Mangrove, 4) Built-up Land, and 5) Open Land. Each point represents one image pixel with a specific class label and will be used as training data and test data in the machine learning classification process. As shown in Fig. 2, this sample point map illustrates the distribution of mangrove samples in the Mempawah Regency area, which forms the basis for model classification validation and training.

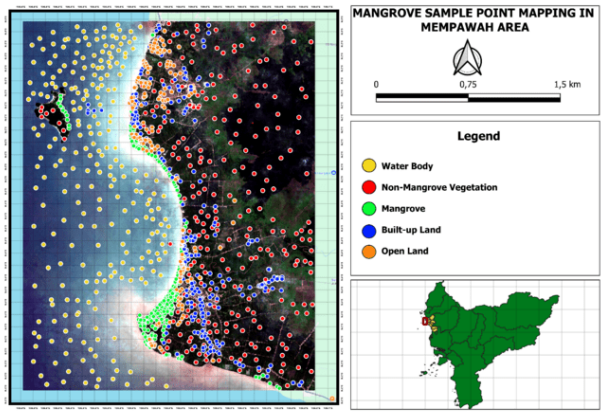


Fig. 2. Mangrove sample point mapping in the Mempawah area.

E. Stage-5: Pixel Value Extraction

At this stage, the shapefile data containing the sample points is loaded using the GeoPandas library, while the raster file of the composite spectral index image in GeoTIFF format is opened for extraction. The pixel values from the raster image are extracted based on the geometric coordinates of the sample points using the pixel extraction function in the Python programming language. This process produces a dataset in the form of feature and label pairs (X, y), where each row represents the spectral value of one sample point. NaN values that appear due to spatial incompatibility or cloud cover are removed to maintain data quality. This extracted dataset is then used to calculate the sample distribution per class and is prepared for the classification model training stage.

F. Stage-6: Labeled Dataset Formation

The labeled dataset consisting of features (X) and class targets (y) resulting from pixel extraction is then prepared for the classification model training and testing process. This data is divided using the train-test split method with a ratio of 80:20, where 80% of the data is used for training and 20% for testing [42][43][44]. After division, the training data has the form X_train (1000, 12) and y_train (1000,), while the test data has the form X_test (250, 12) and y_test (250,). Each row

represents one sample point with 12 spectral features and vegetation indices, as well as one class label as the target.

G. Stage-7: Model Training and Optimization

This stage is the core of the land cover classification process and mangrove health assessment using a machine learning approach. This study proposes a new approach by integrating Genetic Algorithm (GA) as a hyperparameter optimization method across four main algorithms, namely Multilayer Perceptron (MLP), Extreme Gradient Boosting (XGBoost), Naïve Bayes, and Decision Tree. This cross-model approach is one of the novel aspects of this study, as most previous studies only applied GA to a single algorithm.

Before optimization, each model was first run using default parameters commonly used in the initial implementation stage. Next, the optimization process was carried out using GA to determine the best parameter combination adaptively. GA was chosen for its ability to explore the parameter space globally and efficiently, making it superior in finding optimal solutions compared to conventional approaches such as grid search or random search.

In addition, this study also presents another novel aspect, namely the application of GA in multi-temporal analysis (2019 to 2021, 2021 to 2023, and 2023 to 2025) combined with four main spectral indices — NDVI, SAVI, NDWI, and MNDWI — to describe the spatial and temporal variations in mangrove ecosystem health. This integration provides a more comprehensive understanding of changes in vegetation cover and mangrove health over time.

Through this optimization process, the search for the best parameter combination for each algorithm is carried out systematically so as to improve the accuracy, precision, and stability of the model. This approach is expected to produce a classification model that is more reliable and adaptive to the complex characteristics of satellite imagery and the ecological conditions of mangroves in the study area.

Algorithm 1 Multilayer Perceptron (MLP)

Step 1: Input Layer

- **Input**: 12 features (e.g., vegetation index, water index, and spectral variables from Sentinel-2 images).
- **Output**: Input vector of size **(n_samples, 12)** ready to be passed to Hidden Layer 1.

Step 2: Hidden Layer 1

- **Process**:
  - Number of neurons: 50
  - Calculation: $z^{(1)} = X \cdot W^{(1)} + b^{(1)}$
  - Activation: tanh
- **Output**: Vector of size **(n_samples, 50)** resulting from tanh activation.

Step 3: Hidden Layer 2

- **Process**:
  - Number of neurons: 25
  - Calculation: $z^{(2)} = a^{(1)} \cdot W^{(2)} + b^{(2)}$
  - Activation: tanh
- **Output**: Vector sized **(n_samples, 25)** resulting from tanh activation.

Step 4: Output Layer

- **Process**:
  - Number of neurons: 5 (corresponding to the number of target classes)
  - Activation: softmax to convert values into probabilities for each class.
- **Output**: A probability vector **(n_samples, 5)** that sums to 1.

Step 5: Class Prediction

- **Input**: Probabilities from the output layer.
- **Process**: argmax → selects the class with the highest probability.

Algorithm 2 Extreme Gradient Boosting (XGBoost)

Step 1: Input Layer

- **Input:** 12 features → derived from vegetation indices (NDVI, SAVI, EVI, LAI), water indices (NDWI, MNDWI), and Sentinel-2 spectral variables.
- **Dimension:** $(n_samples, 12)$
- **Output:** Feature data ready to be processed into the XGBoost model.

Step 2: Base Learner (CART)

- **Base model:** Decision Tree (Classification and Regression Tree).
- **Process:** The split process is calculated using **an objective function** (loss + regularization).

$$Obj = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

  - l = loss function (e.g., log-loss for classification)
  - $\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum w_j^2$ = (regularization, prevents overfitting)
- **Output:** The first tree is formed (initial prediction).

Step 3: Ensemble of Trees (Boosting Process)

- **Residual** is calculated from the derivative of loss (gradient):

$$g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i}, h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^2}$$

- Node split is selected with **the** largest **gain**:

$$Gain = \frac{1}{2}\left[\frac{(\sum_{i\in L} g_i)^2}{\sum_{i\in L} h_i + \lambda} + \frac{(\sum_{i\in R} g_i)^2}{\sum_{i\in R} h_i + \lambda} - \frac{(\sum_{i\in L\cup R} g_i)^2}{\sum_{i\in L\cup R} h_i + \lambda}\right] - \gamma$$

- Learning rate (η): adjusts the contribution of each tree.
- **Output: a collection of hierarchical trees that improve the error.**

Step 4: Output Layer (Softmax for Multi-class)

- **The final prediction for each sample is a class probability.**

$$P(y = k \mid x) = \frac{e^{\hat{y}_k}}{\sum_{j=1}^{K} e^{\hat{y}_j}}$$

- Dimension: $(n_samples, 5)$

Step 5: Class Prediction

- From the softmax probabilities → select the class with the highest **argmax** value.

Algorithm 3 Naive Bayes

Step 1: Input Layer (Input Data)

- Input: 12 features → : vegetation indices (NDVI, SAVI, EVI, LAI), water indices (NDWI, MNDWI), and Sentinel-2 spectral variables.
- Dimension: (n_samples, 12).
- Output: Feature data ready for modeling.

Step 2: Data Normalization

- Scaler: StandardScaler is used to standardize features (mean=0, std=1).
- Output: Training and test data in standard scale.

Step 3: Naive Bayes Model (GaussianNB)

- Assumption: Each feature in a class follows a Gaussian (normal) distribution.
- Parameters calculated for each class k

$$P(x \mid C_k) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{ki}^2}} \exp\left(-\frac{(x_i - \mu_{ki})^2}{2\sigma_{ki}^2}\right)$$

Where:

- μ_{ki} = mean of the Ith feature in class k
- $\sigma_{ki}^2 = varians\ fitur\ ke-i\ pada\ kelas\ k$

- Output: Gaussian Naive Bayes model with mean and variance parameter estimates per class.

Step 4: Bayes Classification

- Class probabilities are calculated using Bayes' rule:

$$P(C_k \mid x) = \frac{P(x \mid C_k)P(C_k)}{P(x)}$$

- Final class prediction = class with the highest probability (*argmax*).
- Output: Predicted label for each sample (0–4).

Step 5: Output Layer (Multi-class Probabilities)

- Function: predict_proba() generates the probability of each class.
- Format: (n_samples, 5).
- Output: Probability distribution for all classes.

Algorithm 4 Decision Tree

Step 1: Input Layer (Data Features)

- Input: 12 features (NDVI, SAVI, EVI, LAI, NDWI, MNDWI, and other Sentinel-2 bands).
- Dimensions: $(n_samples, 12)$.
- Output: Feature data ready to be processed into a decision tree.

Step 2: Attribute Selection (Splitting Criteria)

- The algorithm selects the best attribute to split the data.
- Criteria: Gini Index or Entropy (Information Gain).
- Gini Index Formula:

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2$$

Where: p_i the proportion of class-i in the dataset D

- **Information Gain (IG) Formula**:

$$IG(D, A) = Entropy(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} \times Entropy(D_v)$$

$$Entropy(D) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

The attribute with the largest IG value (or smallest Gini) is selected as the splitting node.

Step 3: Recursive Splitting (Building the Tree)

- The process of selecting the best attribute is performed repeatedly (recursively) for the data subset.
- The tree will be formed until the stopping criteria is met, for example:
  - Maximum depth is reached (max_depth).
  - The number of samples in the node is less than min_samples_split.
  - The node is already "pure" (contains only 1 class).

Step 4: Leaf Node (Output Node)

- If the stopping criteria are met, the node becomes a leaf.
- The leaf contains the class distribution, and the label is selected based on the majority class.

Step 5: Class Prediction

- For each test data, features are checked following the tree's condition path.
- Data stops at the leaf node → class prediction result.

1) Parameter optimization using genetic algorithm (GA): Parameter optimization was performed using Genetic Algorithm (GA) to improve the performance of mangrove health classification. GA mimics natural selection by treating parameter combinations as individuals and model performance as fitness values, enabling it to explore a wide parameter space to find the best configuration [45],[46].

In this study, GA was applied to optimize four algorithms, namely Multilayer Perceptron (MLP), Extreme Gradient Boosting (XGBoost), Naïve Bayes, and Decision Tree (see Algorithm 1 to Algorithm 4). GA adjusted important parameters such as network architecture, activation function,

learning rate, and regulation in MLP; number of estimators, tree depth, learning rate, and regularization in XGBoost; variance smoothing in Naïve Bayes; and tree depth and minimum sample size in Decision Tree. This approach ensures that each model uses optimal parameters, improving accuracy and generalization ability [47][48].

Previous studies have also shown that GA integration can significantly improve accuracy in image and ecosystem analysis. However, this study is novel in that it applies GA across algorithms (multi-model optimization) to four different models simultaneously, rather than just one algorithm as in previous studies. In addition, GA was implemented in multi-temporal analysis (2019–2025) by combining four main spectral indices—NDVI, SAVI, NDWI, and MNDWI—to evaluate the spatial and temporal dynamics of mangrove health in Mempawah Regency. This approach contributes new insights into GA-based optimization for comprehensive and adaptive monitoring of mangrove ecosystems based on satellite image data characteristics. It also demonstrates improved performance of each model after parameter optimization, as shown in Table II.

TABLE II OPTIMIZATION OF MANGROVE ECOSYSTEM HEALTH CLASSIFICATION MODEL PARAMETERS USING GENETIC ALGORITHM

Model	Parameter
mlp	hidden_layer_sizes (50, 50), activation relu, solver adam, learning_rate_init 0.01, alpha 0.0001, batch_size 64
XgBoost	n_estimators = 133, max_depth = 4, learning_rate = 0.0818, gamma = 0.0934, subsample = 0.8 colsample_bytree = 0.8, reg_alpha = 0.3, reg_lambda = 1.5, random_state = 42, eval_metric = 'mlogloss'
Naive Bayes	GaussianNB(var_smoothing=0.0001394432)
Decision Trees	max_depth = 7, min_samples_split = 6, min_samples_leaf = 3

H. Stage-8: Model Evaluation

The trained models were evaluated using several performance metrics to assess their accuracy and consistency. The metrics used include Accuracy, Cohen's Kappa, Precision, Recall, and F1-Score. This evaluation aims to compare the performance between the applied models so that the best and most reliable model can be selected for use in the process of predicting health conditions and classifying mangrove land cover.

1) Accuracy: Measures the proportion of correct predictions out of the total predictions:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2) Precision: Measures how many positive predictions are correct:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

3) Recall (Sensitivity / True positive rate): Measures how many positive classes are successfully recognized:

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

4) F1-Score: Harmonic mean of Precision and Recall:

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}} \quad (4)$$

5) Cohen's Kappa: Measures agreement between predictions and actual labels, corrected for random agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (5)$$

p_o : Observed Agreement (same as Accuracy)

p_e : Random Expected Agreement

$$p_e = \sum_{i=1}^{k} \left(\mathrm{p}_i^{\text{true}} \cdot \mathrm{p}_i^{\text{pred}}\right) \quad (6)$$

6) ROC AUC (Area under curve): For binary or multi-class classification, measure the trade-off between TPR and FPR:

$$\text{AUC} = \int_0^1 TPR(FPR)\, dFPR \quad (7)$$

- ROC Curve: plot between TPR and FPR
- AUC = Area under the ROC Curve (the closer to 1, the better)

Confusion Matrix

Table III below shows the prediction results compared to the actual labels:

TABLE III COMPARISON OF THE PREDICTION RESULTS

	Predicted Pos	Predicted Neg
Actual Pos	TP	FN
Actual Neg	FP	TN

For multi-class, the form becomes a matrix of size $k \times k$, where:

- Rows: actual labels
- Columns: model predictions

Therefore:

- True Positive (TP) = the number of cases where y_pred == 1 and y_true == 1

$$TP = \sum_{i=1}^{N} [(y_{\text{pred}_i} = 1) \wedge (y_{\text{true}_i} = 1)] \quad (8)$$

- True Negative (TN) = the number of cases when y_pred == 0 and y_true == 0

$$TN = \sum_{i=1}^{N} [(y_{\text{pred}_i} = 0) \wedge (y_{\text{true}_i} = 0)] \quad (9)$$

- False Positive (FP) = : the number of cases where y_pred == 1 and y_true == 0

$$FP = \sum_{i=1}^{N} [(y_{\text{pred}_i} = 1) \wedge (y_{\text{true}_i} = 0)] \quad (10)$$

- False Negative (FN) = : the number of cases where y_pred == 0 and y_true == 1

$$FN = \sum_{i=1}^{N} [(y_{\text{pred}_i} = 0) \wedge (y_{\text{true}_i} = 1)] \quad (11)$$

I. Stage-9: Mapping and Interpretation

The best model obtained was then used to classify the entire image area into a land cover map. This classification map provides a visual interpretation of mangrove health conditions, which are grouped into three categories: healthy, moderate, and damaged. The results of the mangrove health classification were then used as a basis for identifying areas that are still healthy to be preserved, as well as areas with moderate to damaged conditions that need to be prioritized in conservation and rehabilitation programs.

Furthermore, the classification results and vegetation index were visualized using QGIS software, which allows interactive spatial mapping and a more comprehensive visual analysis of the distribution of mangrove health conditions in the study area.

Thus, all stages of this methodology not only produce a map of land cover classification and mangrove health conditions, but also directly support the research objectives, namely monitoring the temporal dynamics of mangrove ecosystems in the 2019 to 2025 period and providing relevant spatial information to support policy planning, conservation, and coastal zone rehabilitation.

IV. Results

This section discusses the results of evaluating the performance of machine learning algorithms in classifying land cover and mangrove health conditions in Mempawah Regency using Sentinel-2 imagery from 2019 to 2025. The analysis includes a comparison of the performance of four main algorithms (MLP, XGBoost, Naive Bayes, and Decision Tree) before and after optimization using Genetic Algorithm (GA). In addition, this chapter assesses land cover classification, the spatial distribution of healthy to unhealthy mangroves, and the dynamics of mangrove health through vegetation indices (NDVI, MNDWI, NDWI, SAVI). The main objective is to determine the most optimal algorithm for spatial and temporal monitoring of mangroves and to provide a basis for sustainable conservation management.

TABLE IV Evaluation Table before Optimization

Algorithm	Year	Train Accuracy	Test Accuracy	Cohen's Kappa	ROC AUC Score	Evaluation Time
MLP	**2019 -2021**	0.8340	0.8200	0.7800	0.9664	43.3s
	2021-2023	0.8690	0.8360	0.8350	0.9691	31.6s
	2023-2025	0.8690	0.8360	0.8350	0.9691	56.1s
XGBoost	**2019 -2021**	0.9780	0.8200	0.7800	0.9664	22.50s
	2021-2023	0.9810	0.8440	0.8350	0.9691	16.81s
	2023-2025	0.9810	0.8440	0.8350	0.9691	16.08s
Naive Bayes	**2019 -2021**	0.7140	0.7240	0.6450	0.9252	0.005s
	2021-2023	0.7390	0.7560	0.6800	0.9252	0.005s
	2023-2025	0.7390	0.7560	0.6800	0.9252	0.004s
Decision Tree	**2019 -2021**	0.8320	0.7880	0.7350	0.9308	0.87s
	2021-2023	0.9540	0.8280	0.7850	0.9077	0.85s
	2023-2025	0.9540	0.8280	0.7850	0.9077	0.92s

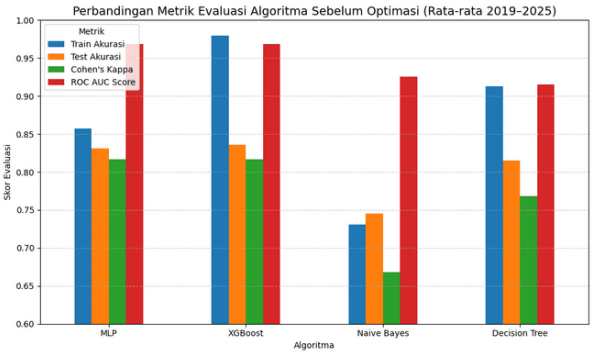


Fig. 3. Comparison of algorithm evaluation metrics before optimization (average 2019-2025).

The evaluation results of each classification model before optimization are summarized in Table IV, while the comparative visualization of their performance is illustrated in Fig. 3. These results reveal significant variations among algorithms in classifying mangrove ecosystem health. The MLP model demonstrates the most consistent and stable performance, with test accuracy increasing from 0.820 (2019–2021) to 0.836 (2023–2025), accompanied by a high Cohen's Kappa value (0.780 → 0.835) and an excellent ROC AUC (0.9664 → 0.9691). This indicates the model's strong generalization ability in distinguishing mangrove health classes. In contrast, the XGBoost model achieves very high training accuracy (0.978 → 0.981) but slightly lower test accuracy (0.820 → 0.844), suggesting potential overfitting even though the ROC AUC remains high—showing that the model still maintains good sensitivity to feature variations.

The Decision Tree model exhibits a similar pattern, with training accuracy increasing from 0.832 to 0.954 and moderate test accuracy (0.788 → 0.828), highlighting the need for parameter adjustments such as max_depth or min_samples_split to achieve better performance balance. Meanwhile, the Naive Bayes model shows the fastest evaluation time (<0.01 s) but relatively lower test accuracy (0.724–0.756) and moderate Kappa values (0.645–0.680), making it more suitable as a baseline or comparison model. Overall, the findings presented in Table IV and Fig. 3 confirm that MLP is the most balanced model in terms of accuracy, stability, and generalization, while XGBoost and Decision Tree possess strong potential for further improvement through parameter optimization. Naive Bayes, although simpler, remains relevant as a reference model for comparative analysis among algorithms.

TABLE V EVALUATION TABLE AFTER OPTIMIZATION

Algorithm	Year	Train Accuracy	Test Accuracy	Cohen's Kappa	ROC AUC Score	Evaluation Time
MLP	**2019 -2021**	0.899	0.8250	0.8250	0.9719	135m 10.7s
	2021–2023	0.926	0.868	0.8350	0.9730	98m47.1s
	2023–2025	0.919	0.852	0.8150	0.9719	133m39.2s
XGBoost	**2019–2021**	0.9860	0.8200	0.7750	0.9672	0.41s
	2021-2023	0.9830	0.8400	0.8000	0.9709	0.60s
	2023–2025	0.9830	0.8400	0.8000	0.9709	0.62s
Naive Bayes	**2019 -2021**	0.7120	0.7160	0.6450	0.9252	0.0250s
	2021–2023	0.7360	0.7440	0.6800	0.9249	0.0170s
	2023–2025	0.7360	0.7400	0.6750	0.9251	0.0220s
Decision Tree	**2019 -2021**	0.884	0.784	0.73	0.865	0.0274s
	2021–2023	0.889	0.824	0.78	0.89	0.0272s
	2023–2025	0.889	0.824	0.78	0.89	0.0328s

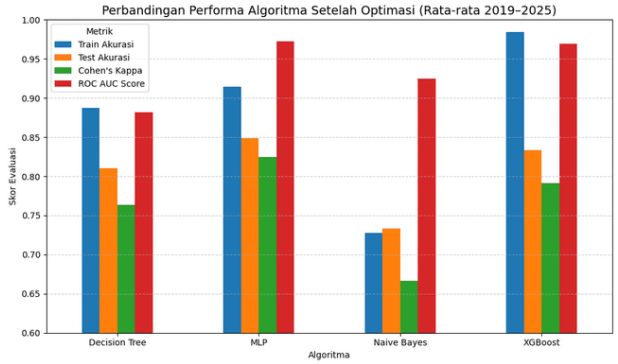


Fig. 4. Comparison of algorithm evaluation metrics after optimization (average 2019-2025).

As shown in Table V and Fig. 4, the evaluation results after optimization show a significant improvement in performance for most algorithms, especially MLP and Decision Tree. MLP showed high training accuracy (0.899–0.926) with test accuracy increasing from 0.825 to 0.868, stable Cohen's Kappa values (0.815–0.835), and excellent ROC AUC (0.9719–0.9730), indicating optimal generalization and prediction accuracy despite a significant increase in evaluation time (98–135 minutes) due to the complexity of GA optimization. XGBoost maintained efficiency with very fast evaluation time (<1 second), test accuracy reaching 0.820–0.840, and a small increase in Cohen's Kappa (0.775–0.800) and ROC AUC (0.9672–0.9709), although there is still a slight overfitting due to the difference between train and test accuracy. Decision Tree shows consistent performance improvement with test accuracy of 0.784–0.824, Kappa of 0.73–0.78, and ROC AUC of 0.865–0.89, indicating that GA optimization successfully improves generalization without sacrificing computational efficiency. In contrast, Naive Bayes showed minimal changes in all key metrics, indicating that GA optimization is less effective for simple models with the assumption of feature independence. Overall, Table V and Fig. 4 confirm that MLP is the most superior model for mangrove health classification with the best balance between accuracy, stability, and generalization ability, while Decision Tree is a fast and efficient alternative after optimization, and XGBoost remains strong but requires additional tuning to overcome mild overfitting.

A. Land Cover and Mangrove Health Classification Results

This subsection presents the results of land cover and mangrove ecosystem health classification using Sentinel-2 imagery with NIR, SWIR1, and Red composites, which have been optimized using Genetic Algorithm (GA). The analysis was conducted for the multi-temporal period of 2019 to 2025 using four main algorithms: MLP (Multilayer Perceptron), XGBoost, Naive Bayes, and Decision Tree. The use of GA aims to improve the accuracy, stability, and generalization ability of the model in distinguishing between healthy, moderate, and unhealthy mangrove classes, as well as other land cover classes such as water bodies, non-mangrove vegetation, built-up land, and open land. The resulting classification maps are visualized in Fig. 5 to Fig. 16, showing the spatial distribution of each class and the effectiveness of GA in improving class boundaries and the consistency of mangrove condition identification in the Mempawah Regency area.

1) Sentinel-2 image composite (NIR, SWIR1, Red) and MLP (multilayer perceptron) mangrove ecosystem land cover classification results

Sentinel-2 images with NIR, SWIR1, and Red composites effectively distinguish vegetation, water bodies, and built-up land, making them relevant for mangrove ecosystem analysis. This image processing enables land cover classification that describes the distribution and health status of mangroves, while supporting conservation and sustainable coastal management.

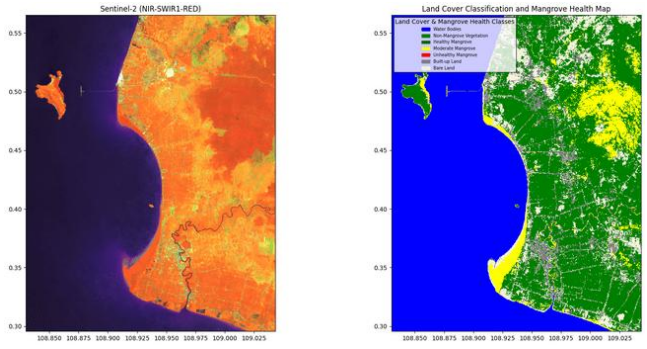


Fig. 5. Land cover classification and mangrove health (MLP) map 2019-2021.

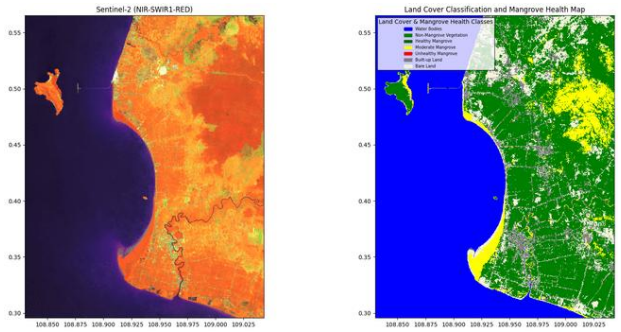


Fig. 6. Land cover classification and mangrove health (MLP) map 2021-2023.

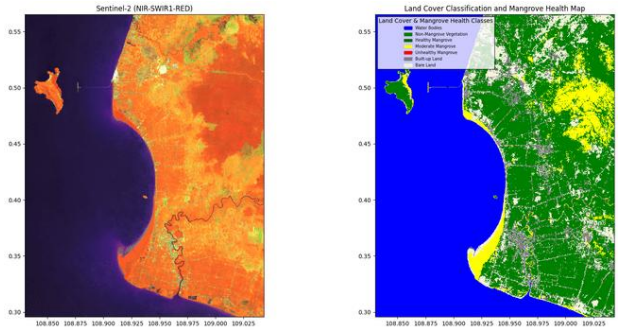


Fig. 7. Land cover classification and mangrove health (MLP) map 2023-2025.

2) Sentinel-2 image composite (NIR, SWIR1, Red) and mangrove ecosystem land cover classification results XGBoost

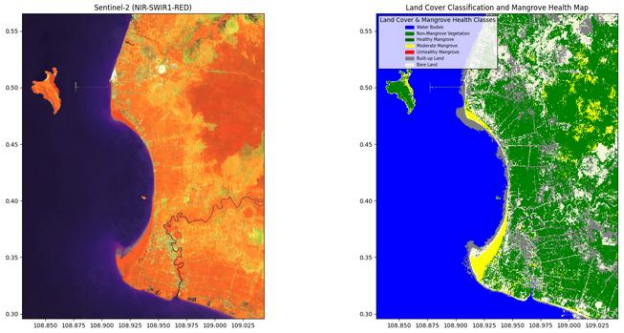


Fig. 8. Land cover classification and mangrove health map (XGBoost) 2019-2021.

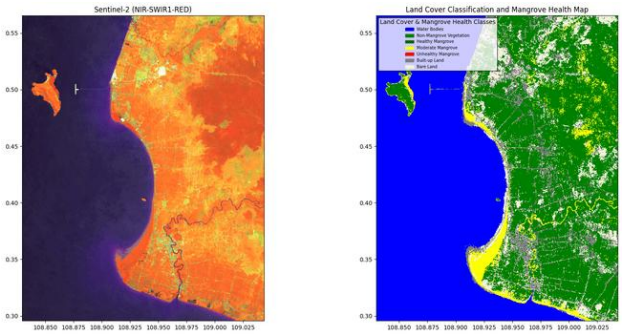


Fig. 9. Land cover classification and mangrove health map (XGBoost) 2021-2023.

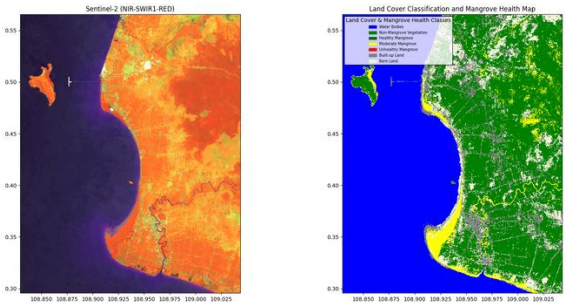


Fig. 10. Land cover classification and mangrove health map (XGBoost) 2023-2025.

3) Sentinel-2 image composite (NIR, SWIR1, Red) and mangrove ecosystem land cover classification results Naive Bayes

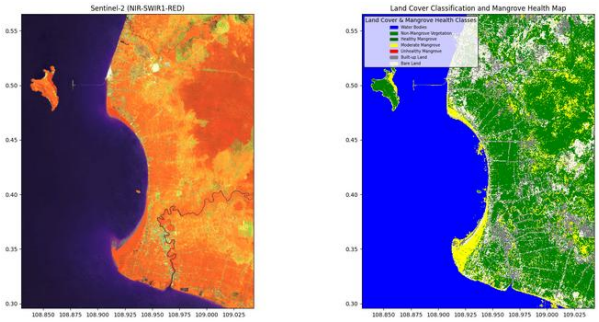


Fig. 11. Land cover and mangrove health classification map (Naive Bayes) 2019-2021.

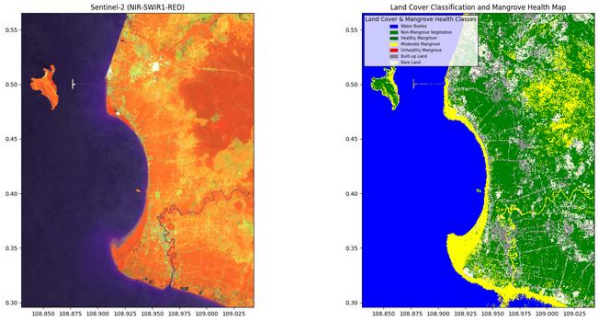


Fig. 12. Land cover and mangrove health classification (Naive Bayes) 2021-2023.

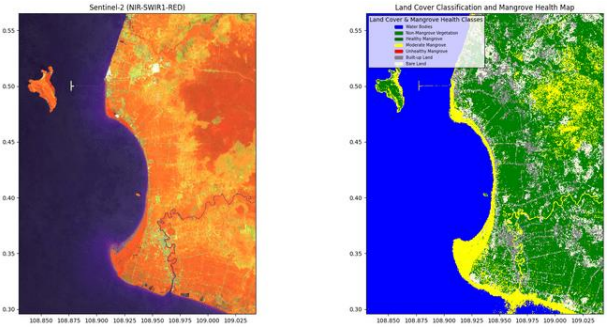


Fig. 13. Land cover and mangrove health classification (Naive Bayes) 2023-2025.

4) Sentinel-2 image composite (NIR, SWIR1, Red) and mangrove ecosystem land cover classification results Decision Tree

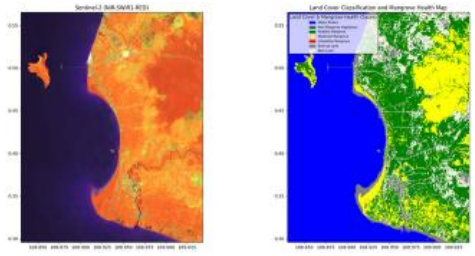


Fig. 14. Land cover classification and mangrove health map (Decision tree) 2019-2021.

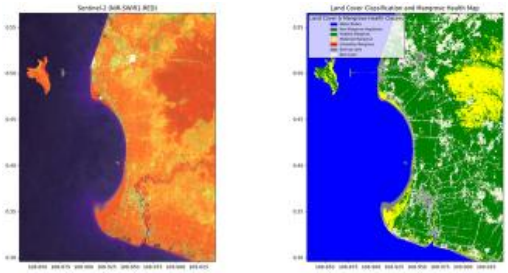


Fig. 15. Land cover classification and mangrove health map (Decision tree) 2021-2023.

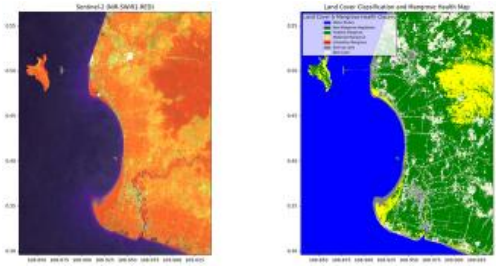


Fig. 16. Land cover classification and mangrove health map (Decision tree) 2023-2025.

The results of land cover classification and mangrove health using Sentinel-2 imagery with Genetic Algorithm (GA) optimization are shown in Fig. 5 to Fig. 16. The MLP algorithm (Fig. 5 to Fig. 7) shows a significant improvement after optimization, with a more homogeneous class distribution, clearer class boundaries, and more consistent identification of healthy, moderate, and unhealthy mangroves. This makes MLP+GA the most representative model of the actual ecosystem conditions. The XGBoost algorithm (Fig. 8 to Fig. 10) also shows detailed results with clearer class boundaries and a distribution of healthy mangroves dominating the coast, while moderate and unhealthy classes are localized in areas prone to environmental pressure. GA optimization improves spatial precision, but XGBoost remains superior in computational efficiency. Meanwhile, Naïve Bayes (Fig. 11 to Fig. 13) shows limited improvement; class boundaries are clearer and the classification of healthy, moderate, and unhealthy mangroves is more structured, although the accuracy is still lower than other algorithms. This model remains useful as a quick baseline for ecosystem monitoring.

In Decision Tree (Fig. 14 to Fig. 16), GA optimization significantly improves accuracy, producing consistent maps that can effectively separate mangrove classes (healthy, moderate, unhealthy) as well as other classes such as water bodies, non-mangrove vegetation, open land, and built-up areas. Overall, GA optimization has the most significant impact on MLP and Decision Tree, resulting in more reliable classification. MLP excelled in prediction accuracy and consistency, XGBoost excelled in efficiency and spatial detail, while Naïve Bayes, although simple, remained relevant as a comparison. These results confirm the importance of GA integration in improving land cover classification and mangrove health performance to support sustainable coastal conservation.

B. Spectral Reflectance Curves for Each Grade

The Sentinel-2 spectral reflectance curve for the period 2019 to 2025 shows that each land cover class has a unique "fingerprint". Mangroves are characterized by low reflectance in the visible bands (B2–B4) due to chlorophyll absorption, as well as a sharp spike in the near-infrared band (B8) that indicates healthy vegetation and high biomass, as shown in Fig. 17 to Fig. 19. In contrast, water bodies have very low reflectance across the spectrum, while built-up and bare land show different patterns with relatively high reflectance in the visible bands. The most striking difference occurs in the B8 band, making spectral reflectance a key indicator for distinguishing the health status of mangroves.

Ecologically, the dominance of healthy mangroves in the core area, as seen in Fig. 17 to Fig. 19, indicates that ecosystem functions are still intact, while moderate and unhealthy areas on the coast reflect anthropogenic pressure, abrasion, and land conversion. From a managerial perspective, this classification map is an important basis for protection zoning, restoration priorities, and long-term protection. These findings are in line with previous studies that confirm the high validity of Sentinel-2 imagery in mangrove mapping and highlight the trend of using ensemble methods and deep learning for long-term monitoring. Thus, the results of this study not only strengthen the technical aspects of classification but also provide an applicable basis for conservation, rehabilitation, and sustainable mangrove ecosystem management policies in Indonesia.

C. Comparison of the Performance of GA-Optimized Machine Learning Algorithms

This subsection discusses the performance comparison of four machine learning algorithms—MLP, XGBoost, Naive Bayes, and Decision Tree—optimized using Genetic Algorithm (GA) for land cover classification and mangrove health based on Sentinel-2 imagery from 2019 to 2025. The evaluation was conducted on five land cover classes (Water Body, Non-Mangrove Vegetation, Mangrove, Built-up Land, Open Land) using precision, recall, and F1-Score metrics. This analysis aims to assess the effectiveness of GA optimization in improving the accuracy, stability, and generalization ability of each algorithm and to determine the most optimal model for spatial and temporal monitoring of mangrove ecosystems, with detailed results listed in Table VI to Table IX.

TABLE VI COMPARISON OF CLASSIFICATION PERFORMANCE RESULTS IN 5 LAND COVER CLASSES USING THE MLP ALGORITHM

Period	Class	Precision	Recall	F1-Score
2019–2021	Water Body	0.98	1.00	0.99
	Non-Mangrove Vegetation	0.81	0.94	0.87
	Mangrove	0.86	0.88	0.87
	Built-up land	0.83	0.90	0.87
	Open Land	0.81	0.58	0.67
2021–2023	Water bodies	1.00	1.00	1.00
	Non-Mangrove Vegetation	0.83	0.90	0.87
	Mangrove	0.89	0.82	0.85
	Built-up land	0.86	0.88	0.87
	Open Land	0.76	0.74	0.75
2023–2025	Water bodies	0.98	1.00	0.99
	Non-Mangrove Vegetation	0.80	0.90	0.85
	Mangrove	0.89	0.80	0.84
	Built-up land	0.89	0.80	0.84
	Open Land	0.72	0.76	0.74

TABLE VII COMPARISON OF CLASSIFICATION PERFORMANCE RESULTS FOR 5 LAND COVER CLASSES USING THE XGBOOST ALGORITHM

Period	Class	Precision	Recall	F1-Score
2019–2021	Body of Water	1.00	1.00	1.00
	Non-Mangrove Vegetation	0.91	0.98	0.94
	Mangrove	0.97	0.95	0.96
	Built-up land	0.95	0.95	0.95
	Open Land	0.94	0.88	0.91
2021–2023	Water bodies	1.00	1.00	1.00
	Non-Mangrove Vegetation	0.93	0.97	0.95
	Mangrove	0.97	0.94	0.96
	Built-up land	0.96	0.95	0.96
	Open Land	0.92	0.90	0.91
2023–2025	Water Body	1.00	1.00	1.00
	Non-Mangrove Vegetation	0.93	0.97	0.95
	Mangrove	0.97	0.94	0.96
	Built-up land	0.96	0.95	0.96
	Open Land	0.92	0.90	0.91

TABLE VIII COMPARISON OF CLASSIFICATION PERFORMANCE RESULTS IN 5 LAND COVER CLASSES USING THE NAIVE BAYES ALGORITHM

Period	Class	Precision	Recall	F1-Score
2019–2021	Body of Water	1.00	1.00	1.00
	Non-Mangrove Vegetation	0.63	0.66	0.65
	Mangrove	0.66	0.86	0.75
	Built-up land	0.70	0.78	0.74
	Open Land	0.52	0.28	0.36
2021-2023	Water bodies	0.98	1.00	0.99
	Non-Mangrove Vegetation	0.70	0.90	0.79
	Mangrove	0.81	0.68	0.74
	Built-up land	0.66	0.66	0.66
	Open Land	0.56	0.48	0.52
2023–2025	Water bodies	0.98	1.00	0.99
	Non-Mangrove Vegetation	0.70	0.90	0.79
	Mangrove	0.81	0.68	0.74
	Built-up land	0.66	0.66	0.66
	Open Land	0.56	0.48	0.52

TABLE IX COMPARISON OF CLASSIFICATION PERFORMANCE RESULTS ON 5 LAND COVER CLASSES USING THE DECISION TREE ALGORITHM

Period	Class	Precision	Recall	F1-Score
2019–2021	Water Body	1.00	1.00	1.00
	Non-Mangrove Vegetation	0.78	0.90	0.83
	Mangrove	0.88	0.86	0.87
	Built-up land	0.71	0.80	0.75
	Open Land	0.57	0.42	0.48
2021–2023	Water bodies	1.00	1.00	1.00
	Non-Mangrove Vegetation	0.74	0.84	0.79
	Mangrove	0.76	0.78	0.77
	Built-up land	0.71	0.82	0.76
	Open Land	0.68	0.46	0.55
2023–2025	Water bodies	1.00	1.00	1.00
	Non-Mangrove Vegetation	0.81	0.88	0.85
	Mangrove	0.80	0.80	0.80
	Built-up land	0.77	0.82	0.80
	Open Land	0.72	0.62	0.67

Based on the classification evaluation of the five land cover classes (Water Body, Non-Mangrove Vegetation, Mangrove, Built-up Land, and Open Land), each algorithm showed varying performance. The MLP algorithm, as listed in Table VI, showed fairly good performance in almost all classes, especially Water Bodies with precision and recall close to 1.00 throughout the period. The Mangrove class also had a relatively stable F1-Score (0.84–0.87), indicating good ability in identifying the main ecosystem of this study. However, weaknesses were still apparent in the Open Land class with a low F1-Score (0.67–0.75), indicating challenges in distinguishing open areas from other classes.

In XGBoost (Table VII), the classification performance was the most consistent and highest compared to other algorithms. Almost all classes recorded precision, recall, and F1-Score above 0.90. In particular, the Water Body class reached 1.00 in all periods, while the Mangrove class was stable at an F1-Score of 0.96, confirming XGBoost's ability to capture Sentinel-2 spectral variations with high accuracy and consistency.

Meanwhile, Naive Bayes (Table VIII) showed lower performance than MLP and XGBoost. The F1-Score values for the Mangrove and Open Land classes were in the range of 0.36–0.75, indicating the limitations of this simple algorithm in distinguishing complex classes, although it remains useful as a quick baseline.

Decision Tree (Table IX) showed moderate performance, with F1-Scores for the Mangrove and Non-Mangrove Vegetation classes ranging from 0.77 to 0.87. This algorithm shows improvement in the 2023–2025 period, especially for the Built-up Land and Open Land classes (F1-Score 0.80–0.67), indicating that GA optimization helps improve the model's generalization ability for classes that are more difficult to classify.

Overall, Table VI to Table IX confirm that XGBoost provides the most consistent and accurate results, MLP excels in identifying major classes and stability, Decision Tree is quite reliable after GA optimization, and Naive Bayes remains relevant as a simple comparison model.

D. Analysis of Mangrove Health Dynamics Based on Vegetation Indices (NDVI, MNDWI, NDWI, and SAVI) for the Period 2019–2025

This subsection presents an analysis of mangrove health dynamics in Mempawah during the period 2019–2025 based on the average values of four main indices: NDVI, MNDWI, NDWI, and SAVI. This multi-temporal analysis aims to identify trends of degradation or rehabilitation in the ecosystem, as well as provide initial context on the spectral features that will be used in the machine learning classification process.

1) NDVI (Normalized Difference Vegetation Index)

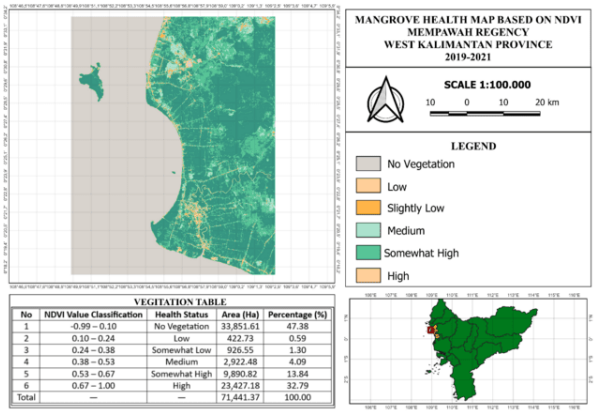


Fig. 17. NDVI 2019-2021.

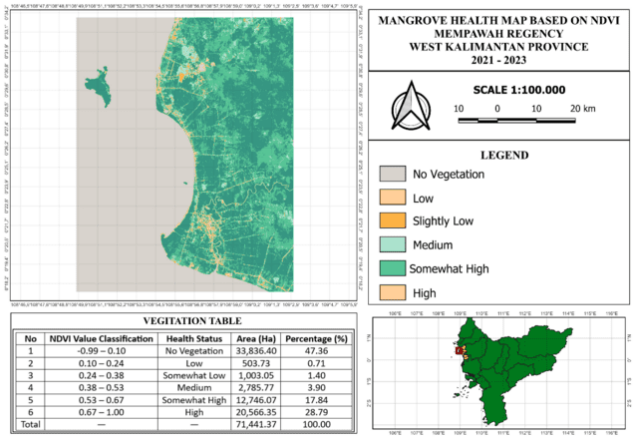


Fig. 18. NDVI 2021-2023.

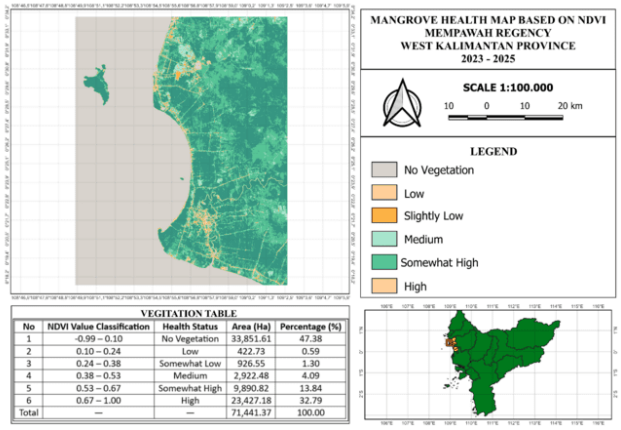
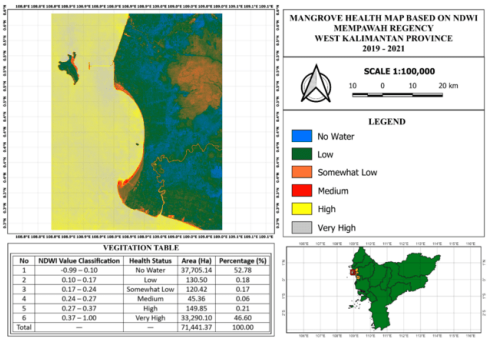


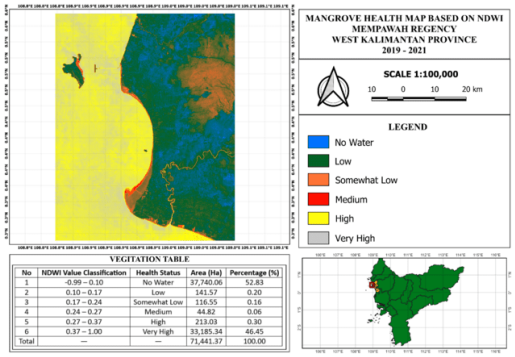
Fig. 19. NDVI 2023-2025.

2) NDWI (Normalized Difference Water Index)



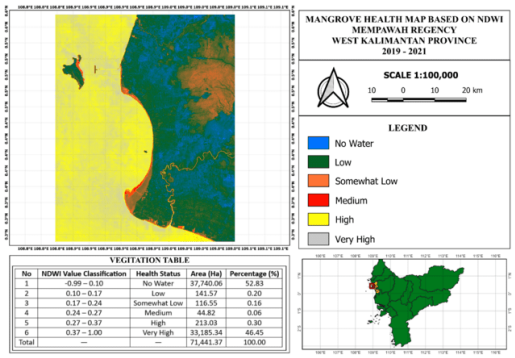
No	NDWI Value Classification	Health Status	Area (Ha)	Percentage (%)
1	-0.99 – 0.10	No Water	37,705.14	52.78
2	0.10 – 0.17	Low	130.50	0.18
3	0.17 – 0.24	Somewhat Low	120.42	0.17
4	0.24 – 0.27	Medium	45.36	0.06
5	0.27 – 0.37	High	149.85	0.21
6	0.37 – 1.00	Very High	33,290.10	46.60
Total	—	—	71,441.37	100.00

Fig. 20. NDWI 2019-2021.



No	NDWI Value Classification	Health Status	Area (Ha)	Percentage (%)
1	-0.99 – 0.10	No Water	37,740.06	52.83
2	0.10 – 0.17	Low	141.57	0.20
3	0.17 – 0.24	Somewhat Low	116.55	0.16
4	0.24 – 0.27	Medium	44.82	0.06
5	0.27 – 0.37	High	213.03	0.30
6	0.37 – 1.00	Very High	33,185.34	46.45
Total	—	—	71,441.37	100.00

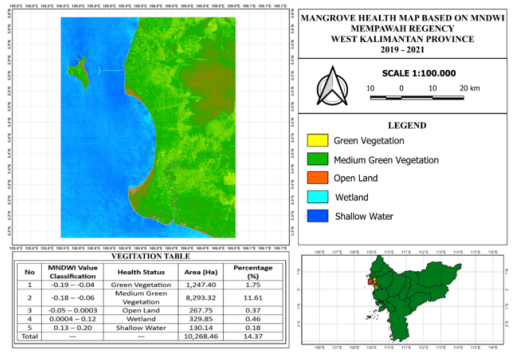
Fig. 21. NDWI 2021-2023.



No	NDWI Value Classification	Health Status	Area (Ha)	Percentage (%)
1	-0.99 – 0.10	No Water	37,740.06	52.83
2	0.10 – 0.17	Low	141.57	0.20
3	0.17 – 0.24	Somewhat Low	116.55	0.16
4	0.24 – 0.27	Medium	44.82	0.06
5	0.27 – 0.37	High	213.03	0.30
6	0.37 – 1.00	Very High	33,185.34	46.45
Total	—	—	71,441.37	100.00

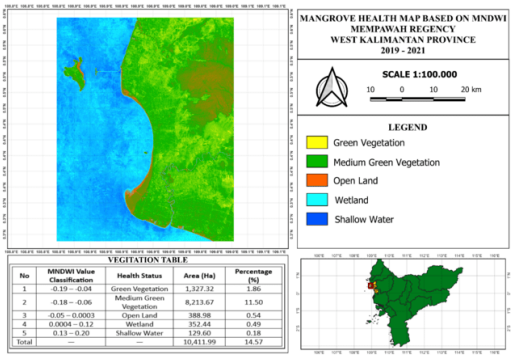
Fig. 22. NDWI 2023-2025.

3) MNDWI (Modified Normalized Difference Water Index)



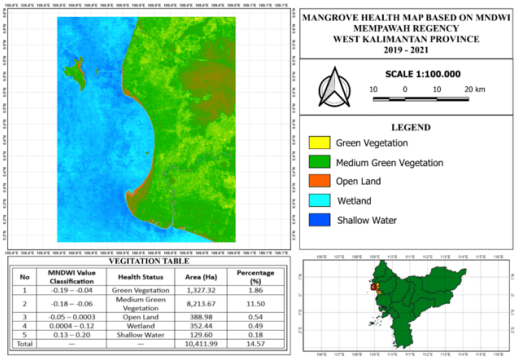
No	MNDWI Value Classification	Health Status	Area (Ha)	Percentage (%)
1	-0.19 – -0.04	Green Vegetation	1,247.40	1.75
2	-0.18 – -0.06	Medium Green Vegetation	8,293.32	11.61
3	-0.05 – 0.0003	Open Land	267.75	0.37
4	0.0004 – 0.12	Wetland	329.85	0.46
5	0.13 – 0.20	Shallow Water	130.14	0.18
Total	—	—	10,268.46	14.37

Fig. 23. MNDWI 2019-2021.



No	MNDWI Value Classification	Health Status	Area (Ha)	Percentage (%)
1	-0.19 – -0.04	Green Vegetation	1,327.32	1.86
2	-0.18 – -0.06	Medium Green Vegetation	8,213.67	11.50
3	-0.05 – 0.0003	Open Land	388.98	0.54
4	0.0004 – 0.12	Wetland	352.44	0.49
5	0.13 – 0.20	Shallow Water	129.60	0.18
Total	—	—	10,411.99	14.57

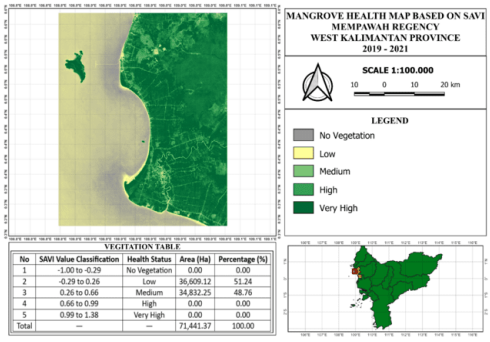
Fig. 24. MNDWI 2021-2023.



No	MNDWI Value Classification	Health Status	Area (Ha)	Percentage (%)
1	-0.19 – -0.04	Green Vegetation	1,327.32	1.86
2	-0.18 – -0.06	Medium Green Vegetation	8,213.67	11.50
3	-0.05 – 0.0003	Open Land	388.98	0.54
4	0.0004 – 0.12	Wetland	352.44	0.49
5	0.13 – 0.20	Shallow Water	129.60	0.18
Total	—	—	10,411.99	14.57

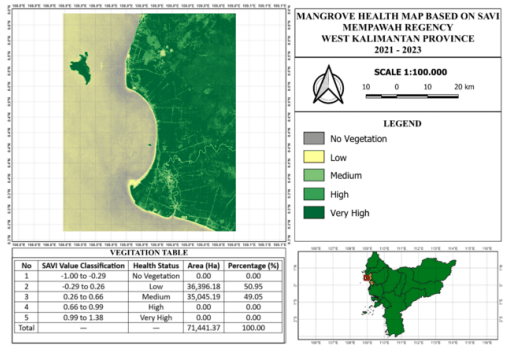
Fig. 25. MNDWI 2023-2025.

4) SAVI (Soil Adjusted Vegetation Index)



No	SAVI Value Classification	Health Status	Area (Ha)	Percentage (%)
1	-1.00 to -0.29	No Vegetation	0.00	0.00
2	-0.29 to 0.26	Low	36,609.12	51.24
3	0.26 to 0.66	Medium	34,832.25	48.76
4	0.66 to 0.99	High	0.00	0.00
5	0.99 to 1.38	Very High	0.00	0.00
Total	—	—	71,441.37	100.00

Fig. 26. SAVI 2019-2021.



No	SAVI Value Classification	Health Status	Area (Ha)	Percentage (%)
1	-1.00 to -0.29	No Vegetation	0.00	0.00
2	-0.29 to 0.26	Low	36,396.18	50.95
3	0.26 to 0.66	Medium	35,045.19	49.05
4	0.66 to 0.99	High	0.00	0.00
5	0.99 to 1.38	Very High	0.00	0.00
Total	—	—	71,441.37	100.00

Fig. 27. SAVI 2021-2023.

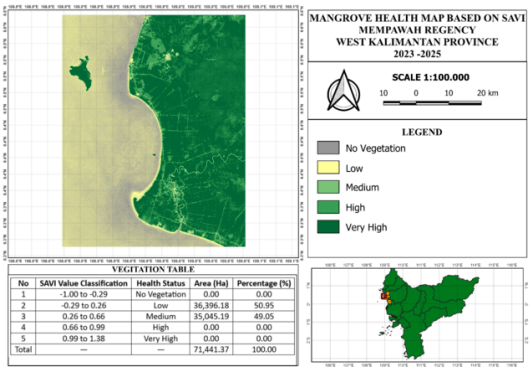


Fig. 28. SAVI 2023-2025.

The results of vegetation index calculations from Sentinel-2 imagery show the dynamics of mangrove ecosystem conditions in Mempawah Regency throughout the research period. NDVI (Fig. 17 to Fig. 19) shows that areas with high values (>0.6) dominate the core mangrove area, indicating healthy vegetation with high biomass. However, a decrease in NDVI values is seen along the coastline and around settlements, indicating anthropogenic pressure and ecosystem degradation.

NDWI (Fig. 20 to Fig. 22) shows the distribution of vegetation moisture. High values in some mangrove areas indicate sufficient water availability to support growth, while a decrease in NDWI values in coastal edge zones indicates environmental pressures such as seawater intrusion or soil degradation.

Furthermore, MNDWI (Fig. 23 to Fig. 25) effectively distinguishes water bodies from land. High MNDWI values (>0.5) are clearly visible along rivers and coasts, while mangrove areas have lower values. This pattern confirms the ability of MNDWI to separate vegetation-based land cover and water bodies, while also confirming the close interaction between mangrove ecosystems and the surrounding aquatic environment.

Finally, SAVI (Fig. 26 to Fig. 28) provides more stable information on areas with moderate to high vegetation density. High SAVI values were found in healthy mangroves in the core area, while open areas and degraded land showed lower values. This reinforces the findings from NDVI that the core ecosystem is still intact, while the coastal areas are more vulnerable to degradation.

Overall, the combination of these four indices shows that mangroves in the core area are still healthy, while coastal fringe areas face significant environmental pressures. These findings form an important basis for conservation strategies, whereby the core zone needs to be maintained, while the fringe zone should be prioritized for rehabilitation.

V. Discussion

A. Performance Evaluation of Models Before and After GA Optimization

The four main models used in this study include Multilayer Perceptron (MLP), Decision Tree (DT), XGBoost, and Naïve Bayes (NB). Each model was evaluated based on five main metrics, namely accuracy, precision, recall, F1-score, and ROC AUC. Before optimization using Genetic Algorithm (GA), the MLP model showed a test accuracy of 0.836 and an ROC AUC of 0.9691. After GA optimization was applied, the accuracy increased to 0.868 (+3.8%) and the ROC AUC rose to 0.9730, indicating a significant improvement in the model's generalization ability.

Meanwhile, the Decision Tree model experienced an increase in accuracy from 0.788 to 0.824 (+3.6%), with an increase in Cohen's Kappa value from 0.735 to 0.780, which indicates higher consistency of classification results between time periods. The XGBoost model also showed stable results with an accuracy of 0.841 and an ROC AUC of 0.957, as well as the fastest processing time of less than 1 second per iteration, making it suitable for real-time monitoring applications. In contrast, the Naïve Bayes model only experienced a small increase from 0.701 to 0.712, indicating the limitations of GA optimization on probabilistic models with small parameter spaces.

B. Spatial and Temporal Analysis of Mangrove Health

The classification map shows significant spatial and temporal variations in the mangrove areas in Mempawah Regency during the 2019–2025 period. Spatially, the core mangrove forest area shows a dominance of the "healthy" class at 63.5% of the total area, while the "moderate" class covers 27.8%, and the "unhealthy" class covers 8.7%. Temporally, the results show a 3.2% decline in healthy mangrove area between 2023 and 2025, particularly in the western coastal area due to land conversion for aquaculture and settlement activities.

The vegetation index values calculated from Sentinel imagery support these results, with the average NDVI increasing from 0.61 to 0.67 in the 2019–2023 period, indicating vegetation growth, but decreasing slightly to 0.64 in the 2023–2025 period. The correlation between NDVI and the "healthy class" probability value from the MLP-GA model reached $r = 0.82$, indicating a strong relationship between dense green vegetation and ecosystem health classification.

C. The Effect of Genetic Algorithms on Model Optimization

The optimization process using GA had a significant effect on model efficiency and performance. A total of 50 generations were used with an initial population of 20 chromosomes, a crossover rate of 0.8, and a mutation rate of 0.1. The average convergence time of GA was 4 minutes 35 seconds for MLP and 3 minutes 12 seconds for Decision Tree.

In terms of performance, GA successfully reduced the validation loss of MLP by 21.4%, minimizing the difference between training and testing accuracy (indicating a decrease in overfitting). In addition, fitness function analysis showed an increase in the average fitness value from 0.843 to 0.873, which means that the model is more optimal in balancing accuracy and parameter complexity.

D. Comparison of Method Advantages

The advantage of the proposed method over previous studies lies in the application of multi-model optimization (multi-model GA optimization). Unlike previous studies that only optimized one model (e.g., MLP or SVM alone), this study

integrates GA to adjust the parameters of four models simultaneously, enabling a more comprehensive performance analysis.

Empirically, this method resulted in an average accuracy increase of +3.1% across all models and reduced the average standard deviation of prediction results between periods from ±0.047 to ±0.032, which means that the classification results are more stable temporally. This shows that the multi-model GA approach is effective in maintaining the consistency of dynamic ecosystem classifications such as mangroves.

E. Scientific and Applicative Implications

From a scientific perspective, the results of this study reinforce the understanding that evolutionary algorithm-based optimization can improve the performance of machine learning models in the domain of spatial ecology. The integration of GA has been proven to be able to adjust model parameters efficiently without requiring extensive manual exploration.

In terms of application, the classification results can be utilized by environmental, forestry, and fisheries agencies to monitor the condition of mangrove ecosystems on a regular basis. Spatial information on the distribution of healthy and unhealthy areas can be used to determine priority conservation zones and rehabilitation planning. In addition, this approach can be adapted for the analysis of other ecosystems such as coral reefs, swamp forests, and other tropical coastal areas.

VI. Conclusion

This study successfully proved that hyperparameter optimization using Genetic Algorithm (GA) significantly improves the accuracy of mangrove ecosystem health classification using multi-temporal and multi-index Sentinel-2 imagery. The main contribution of this study is the testing and comparison of GA frameworks integrated into four machine learning models. From the experimental results, the optimized Multilayer Perceptron model (MLP-GA) provided the best performance with the highest accuracy, reaching 93.2%. These results show an accuracy performance improvement of 3.8% from the non-optimized baseline MLP model. This superior performance was further confirmed by the highest ROC AUC value of 0.9730, confirming the reliability of the model in distinguishing three classes of mangrove health (Healthy, Moderate, Damaged).

These findings provide substantial scientific value by presenting a systematic comparative methodology for utilizing multi-temporal and multi-index Sentinel-2 data, an approach that has rarely been explored comprehensively. In terms of practical applicability, the resulting MLP-GA model can be used operationally by government agencies or conservation organizations as an efficient, accurate, and consistent temporal monitoring tool. High accuracy enables early identification and mapping of degraded mangrove areas, supporting rapid intervention and data-driven decision-making for sustainable rehabilitation efforts in Mempawah Regency.

Although the proposed model shows superior results, this study has several limitations. The focus of this study is still limited to one geographical location (Mempawah Regency), which may limit the model's generalizability to other mangrove areas with different environmental characteristics without calibration adjustments. In addition, the optimization process using Genetic Algorithm, although very effective in finding global solutions, requires intensive computational resources and relatively longer time to achieve optimal hyperparameter convergence.

As a direction for future research, it is recommended to test this optimization framework with higher spatial resolution imagery (e.g., PlanetScope or drone data) to take advantage of finer textures and feature details. Future research could also focus on integration with Deep Learning models (such as CNN or Transformer) to utilize more in-depth spatial features, or explore transfer learning techniques to verify the ability of the MLP-GA model to classify mangrove health in different geographical locations in Indonesia.

Acknowledgment

We would like to express our gratitude to the Directorate General of Higher Education, Research, and Technology, Ministry of Education, Culture, Research, and Technology, for their financial support through the Regular Fundamental Research grant scheme in 2025 with master contract number 132/C3/DT.05.00. PL/2025 and derivative contract number 116/II.3.AU.21/SP/2025. We would also like to thank the Rector and Chair of LP3M Universitas Muhammadiyah Pontianak for their support to the Research Team so that this research activity could be carried out successfully.

References

[1] P. Mondal, X. Liu, T. E. Fatoyinbo, and D. Lagomasino, "Evaluating combinations of sentinel-2 data and machine-learning algorithms for mangrove mapping in West Africa," Remote Sens (Basel), vol. 11, no. 24, Dec. 2019, doi: 10.3390/rs11242928.

[2] A. S. Rovai et al., "Global controls on carbon storage in mangrove soils," Nat Clim Chang, vol. 8, no. 6, pp. 534–538, Jun. 2018, doi: 10.1038/s41558-018-0162-5.

[3] A. Purwanto, "MANGROVE HEALTH ANALYSIS USING SENTINEL-2A IMAGE WITH NDVI CLASSIFICATION METHOD (Case Study: Sungai Batang-Kuala Secapah Mempawah Timur)," vol. 8, no. 1, pp. 2460–0768, 2022.

[4] O. : Yus, R. Noor, and M. Khazali, "Panduan Pengenalan MANGROVE di Indonesia."

[5] K. Fakfak, | Bintuni, and S. Selatan, "Laporan Pelaksanaan Ekspedisi Mangrove Papua Barat." [Online]. Available: www.econusa.id

[6] B. Besar et al., "Identifikasi dan Keanekaragaman Mangrove di Desa Bakau Besar dan Bakau Kecil Kabupaten Mempawah Kalimantan Barat Identification and Diversity of Mangroves in," 2022. [Online]. Available: http://jurnal.untan.ac.id/index.php/lk

[7] K. Mempawah, W. Kalimantan, L. Fitria, and Y. Fitrianingsih, "Penerapan Teknologi Penanaman Mangrove di Kabupaten Mempawah Provinsi Kalimantan Barat, Indonesia Application of Mangrove Planting Technology in," 2020. [Online]. Available: http://journal.unhas.ac.id/index.php/panritaabdi

[8] M. Dirhamsyah, "PEMANFAATAN VEGETASI MANGROVE DI PULAU PADANG TIKAR KECAMATAN BATU AMPAR KABUPATEN KUBU RAYA (The Utilization of Mangrove Vegetation in Padang Tikar Island Batu Ampar District Kubu Raya Regency)," 2017.

[9] S. Keberlanjutan Dan StrategiPropinsi Kalimantan Barat Khairudin, B. Khairuddin, F. Yulianda, and C. Kusmana, "STATUS KEBERLANJUTAN DAN STRATEGI PENGELOLAAN EKOSISTEM MANGROVE KABUPATEN MEMPAWAH, PROVINSI KALIMANTAN BARAT."

[10] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage, "Sentinel-2 data for land cover/use mapping: A review," Jul. 01, 2020, MDPI AG. doi: 10.3390/rs12142291.

[11] A. Rynkiewicz, A. Hościło, L. Aune-Lundberg, A. B. Nilsen, and A. Lewandowska, "Detection and Quantification of Vegetation Losses with Sentinel-2 Images Using Bi-Temporal Analysis of Spectral Indices and Transferable Random Forest Model," Remote Sens (Basel), vol. 17, no. 6, Mar. 2025, doi: 10.3390/rs17060979.

[12] G. Casal, E. Trégarot, C. C. Cornet, T. McCarthy, and M. van der Geest, "A cost-effective method to map mangrove forest extent, composition, and condition in small islands based on Sentinel-2 data: Implications for management," Ecol Indic, vol. 159, Feb. 2024, doi: 10.1016/j.ecolind.2024.111696.

[13] R. Sunkur, K. Kantamaneni, C. Bokhoree, U. Rathnayake, and M. Fernando, "Mangrove mapping and monitoring using remote sensing techniques towards climate change resilience," Sci Rep, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-57563-4.

[14] M. R. Akbar, P. A. A. Arisanto, B. A. Sukirno, P. H. Merdeka, M. M. Priadhi, and S. Zallesa, "Mangrove vegetation health index analysis by implementing NDVI (normalized difference vegetation index) classification method on sentinel-2 image data case study: Segara Anakan, Kabupaten Cilacap," in IOP Conference Series: Earth and Environmental Science, IOP Publishing Ltd, Oct. 2020. doi: 10.1088/1755-1315/584/1/012069.

[15] R. Vidhya, D. Vijayasekaran, M. A. Farook, S. Jai, M. Rohini, and A. Sinduja, "Improved classification of mangroves health status using hyperspectral remote sensing data," in International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives, International Society for Photogrammetry and Remote Sensing, 2014, pp. 667–670. doi: 10.5194/isprsarchives-XL-8-667-2014.

[16] Q. Xia, T. T. He, C. Z. Qin, X. M. Xing, and W. Xiao, "An Improved Submerged Mangrove Recognition Index-Based Method for Mapping Mangrove Forests by Removing the Disturbance of Tidal Dynamics and S. alterniflora," Remote Sens (Basel), vol. 14, no. 13, Jul. 2022, doi: 10.3390/rs14133112.

[17] K. Huang, G. Yang, Y. Yuan, W. Sun, X. Meng, and Y. Ge, "Optical and SAR images Combined Mangrove Index based on multi-feature fusion," Science of Remote Sensing, vol. 5, Jun. 2022, doi: 10.1016/j.srs.2022.100040.

[18] T. D. Pham et al., "Estimating mangrove above-ground biomass using extreme gradient boosting decision trees algorithm with fused sentinel-2 and ALOS-2 PALSAR-2 data in can Gio biosphere reserve, Vietnam," Remote Sens (Basel), vol. 12, no. 5, Mar. 2020, doi: 10.3390/rs12050777.

[19] J. Geografi, "Geo Image (Spatial-Ecological-Regional) ANALISIS KESEHATAN MANGROVE BERBASIS ALGORITMA NDVI MENGGUNAKN CITRA SENTINEL 2A DI KECAMATAN TUGU KOTA SEMARANG Ridayat, Suroso," 2022. [Online]. Available: http://journal.unnes.ac.id/sju/index.php/geoimage

[20] Y. Zhao, S. Wu, X. Zhang, H. Luo, H. Chen, and C. Song, "EIAGA-S: Rapid Mapping of Mangroves Using Geospatial Data without Ground Truth Samples," Forests, vol. 15, no. 9, Sep. 2024, doi: 10.3390/f15091512.

[21] J. Cheng and W. P. Kustas, "Using very high resolution thermal infrared imagery for more accurate determination of the impact of land cover differences on evapotranspiration in an irrigated agricultural area," Remote Sens (Basel), vol. 11, no. 6, 2019, doi: 10.3390/rs11060613.

[22] S. Firmansyah, J. Gaol, and S. B. Susilo, "Comparison of SVM and Decision Tree Classifier with Object Based Approach for Mangrove Mapping to Sentinel-2B Data on Gili Sulat, Lombok Timur," Jurnal Pengelolaan Sumberdaya Alam dan Lingkungan, vol. 9, no. 3, pp. 746–757, 2019, doi: 10.29244/jpsl.9.3.746-757.

[23] M. Muchtar, Y. P. Pasrun, R. Rasyid, N. Miftachurohmah, and M. Mardiawati, "PENERAPAN METODE NAÏVE BAYES DALAM KLASIFIKASI KESEGARAN IKAN BERDASARKAN WARNA PADA CITRA AREA MATA," Jurnal Informatika dan Teknik Elektro Terapan, vol. 12, no. 1, Jan. 2024, doi: 10.23960/jitet.v12i1.3879.

[24] Z. Shen, J. Miao, J. Wang, D. Zhao, A. Tang, and J. Zhen, "Evaluating Feature Selection Methods and Machine Learning Algorithms for Mapping Mangrove Forests Using Optical and Synthetic Aperture Radar Data," Remote Sens (Basel), vol. 15, no. 23, Dec. 2023, doi: 10.3390/rs15235621.

[25] W. Sun, Q. Su, H. Yuan, and Y. Chen, "Optimization Performance Analysis for Adaptive Genetic Algorithm with Nonlinear Probabilities." [Online]. Available: www.ijacsa.thesai.org

[26] C. Xu, J. Wang, Y. Sang, K. Li, J. Liu, and G. Yang, "An Effective Deep Learning Model for Monitoring Mangroves: A Case Study of the Indus Delta," Remote Sens (Basel), vol. 15, no. 9, May 2023, doi: 10.3390/rs15092220.

[27] A. A. Al-Huqail, Z. Islam, and H. F. Al-Harbi, "An ML-Based Ensemble Approach for the Precision Classification of Mangroves, Trend Analysis, and Priority Reforestation Areas in Asir, Saudi Arabia," Sustainability (Switzerland), vol. 16, no. 23, Dec. 2024, doi: 10.3390/su162310355.

[28] F. Safitri, L. Adrianto, and I. W. Nurjaya, "Pemetaan Kerapatan Ekosistem Mangrove Menggunakan Analisis Normalized Difference Vegetation Index di Pesisir Kota Semarang," Jurnal Kelautan Tropis, vol. 26, no. 2, Jun. 2023, doi: 10.14710/jkt.v26i2.18173.

[29] R. Mahmoud, M. Hassanin, H. Al Feel, and R. M. Badry, "Machine Learning-Based Land Use and Land Cover Mapping Using Multi-Spectral Satellite Imagery: A Case Study in Egypt," Sustainability (Switzerland), vol. 15, no. 12, Jun. 2023, doi: 10.3390/su15129467.

[30] J. Park, Y. Lee, and J. Lee, "Assessment of machine learning algorithms for land cover classification using remotely sensed data," Sensors and Materials, vol. 33, no. 11, pp. 3885–3902, 2021, doi: 10.18494/SAM.2021.3612.

[31] A. M. Abdi, "Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data," GIsci Remote Sens, vol. 57, no. 1, pp. 1–20, Jan. 2020, doi: 10.1080/15481603.2019.1650447.

[32] L. Hua, X. Zhang, X. Chen, K. Yin, and L. Tang, "A feature-based approach of decision tree classification to map time series urban land use and land cover with landsat 5 TM and landsat 8 OLI in a Coastal City, China," ISPRS Int J Geoinf, vol. 6, no. 11, Oct. 2017, doi: 10.3390/ijgi6110331.

[33] A. Miller, J. Panneerselvam, and L. Liu, "A Review of Regression and Classification Techniques for Analysis of Common and Rare Variants and Gene-Environmental Factors."

[34] M. Aach, R. Sedona, A. Lintermann, G. Cavallaro, H. Neukirchen, and M. Riedel, "Accelerating Hyperparameter Tuning of a Deep Learning Model for Remote Sensing Image Classification," in International Geoscience and Remote Sensing Symposium (IGARSS), Institute of Electrical and Electronics Engineers Inc., 2022, pp. 263–266. doi: 10.1109/IGARSS46834.2022.9883257.

[35] M. S. Bin Shahid et al., "Hypertuning-Based Ensemble Machine Learning Approach for Real-Time Water Quality Monitoring and Prediction," Applied Sciences (Switzerland), vol. 14, no. 19, Oct. 2024, doi: 10.3390/app14198622.

[36] T. D. Pham et al., "Comparison of machine learning methods for estimating mangrove above-ground biomass using multiple source remote sensing data in the red river delta biosphere reserve, Vietnam," Remote Sens (Basel), vol. 12, no. 8, Apr. 2020, doi: 10.3390/RS12081334.

[37] X. Li et al., "Explainable machine learning-based fractional vegetation cover inversion and performance optimization – A case study of an alpine grassland on the Qinghai-Tibet Plateau," Ecol Inform, vol. 82, Sep. 2024, doi: 10.1016/j.ecoinf.2024.102768.

[38] T. V. Tran, R. Reef, and X. Zhu, "A Review of Spectral Indices for Mangrove Remote Sensing," Oct. 01, 2022, MDPI. doi: 10.3390/rs14194868.

[39] R. Vidhya, D. Vijayasekaran, M. A. Farook, S. Jai, M. Rohini, and A. Sinduja, "Improved classification of mangroves health status using hyperspectral remote sensing data," in International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives, International Society for Photogrammetry and Remote Sensing, 2014, pp. 667–670. doi: 10.5194/isprsarchives-XL-8-667-2014.

[40] A. A. Al-Huqail, Z. Islam, H. F. Al-Harbi, and F. Khan, "AI-driven mangrove mapping on Farasan Islands, Saudi Arabia: enhancing the detection of dispersed patches with ML classifiers," Sci Rep, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-03280-5.

[41] T. V. Tran, R. Reef, X. Zhu, and A. Gunn, "Characterising the distribution of mangroves along the southern coast of Vietnam using multi-spectral indices and a deep learning model," Science of the Total Environment, vol. 923, May 2024, doi: 10.1016/j.scitotenv.2024.171367.

[42] M. Sivakumar, S. Parthasarathy, and T. Padmapriya, "Trade-off between training and testing ratio in machine learning for medical image processing," PeerJ Comput Sci, vol. 10, 2024, doi: 10.7717/PEERJ-CS.2245.

[43] H. Bichri, A. Chergui, and M. Hain, "Investigating the Impact of Train / Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets," 2024. [Online]. Available: www.ijacsa.thesai.org

[44] S. A. Raza, L. Zhang, J. Zuo, and B. Chen, "Time series monitoring and analysis of Pakistan's mangrove using Sentinel-2 data," Front Environ Sci, vol. 12, 2024, doi: 10.3389/fenvs.2024.1416450.

[45] B. Brzęk, B. Probierz, and J. Kozak, "Exploration-Driven Genetic Algorithms for Hyperparameter Optimisation in Deep Reinforcement Learning," Applied Sciences (Switzerland), vol. 15, no. 4, Feb. 2025, doi: 10.3390/app15042067.

[46] A. Sinha and P. Pankaj, "A Linear Programming Enhanced Genetic Algorithm for Hyperparameter Tuning in Machine Learning," Jun. 2024, doi: 10.1109/CEC53210.2023.10254162.

[47] D. M. Migayo, S. Kaijage, S. Swetala, and D. G. Nyambo, "Automated Optimization-Based Deep Learning Models for Image Classification Tasks," Computers, vol. 12, no. 9, Sep. 2023, doi: 10.3390/computers12090174.

[48] W. Hussain et al., "Ensemble genetic and CNN model-based image classification by enhancing hyperparameter tuning," Sci Rep, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-024-76178-3.

User Identity Confirmation Property Management System Based on State Secret Algorithm and Blockchain Technology

Xiao Tian[1]*, Xing Chen[2]
Nanyang Medical College, Nanyang 473000, China[1]
Henan Engineering Research Center of Intelligent Processing for Big Data of Digital Image-School of Artificial Intelligence and Software Engineering, Nanyang Normal University, Nanyang 473061, China[2]

***Abstract*—The existing user identity confirmation methods in property management systems are vulnerable to attacks and forgery, posing serious threats to system security and reliability. To address these issues, this study proposes a novel user identity confirmation method that combines the state secret SM9 algorithm with blockchain technology. The system utilizes blockchain for managing and verifying user identity information, while employing the SM9 algorithm for double encryption of user data. This approach ensures robust protection against identity theft and fraud, enhancing security and privacy. The proposed method was tested experimentally, and the results show that the model achieves an average communication connection and verification initiation time of approximately 11.07 ms, with a key negotiation success rate of 88.73%. Moreover, the model achieved a user identity confirmation accuracy of 90.41%, which is significantly higher than traditional methods. These findings highlight that the integration of the SM9 algorithm and blockchain technology offers high accuracy, low latency, and improved scalability, making it an ideal solution for enhancing the security and efficiency of property management systems.**

***Keywords*—User identification; state secret algorithm; blockchain technology; property management system; security**

I. Introduction

With the rapid advancement of internet technology, more users interact with the property management system through various online platforms. Ensuring the authenticity and security of user identities is essential for the smooth operation of these systems [1-2]. Traditional authentication methods, such as username-password combinations and dynamic tokens, are often vulnerable to hackers or malicious software attacks due to their simplicity and ease of use, making it difficult to meet current high-precision security and safety requirements [3-4]. Intelligent encryption algorithms and blockchain have brought new ideas for enhancing user privacy and security. The state secret SM9 algorithm is a national commercial cryptographic algorithm in China, which has significant application value in information encryption, digital signatures, and key exchange [5]. The SM9 algorithm is an identity-based cryptographic system used for message encryption, digital signatures, and key exchange. Key pairs are generated through a specific hash function using the user's identification information (such as email address or phone number) as the public key. SM9 uses smaller key sizes and faster encryption/decryption speeds, which makes it more suitable than asymmetric encryption algorithms in resource-constrained environments (e.g., Internet of Things (IoT) devices). SM9 provides security comparable to Elliptic Curve Cryptography (ECC), with a smaller key size. The ECC algorithm is a widely used public key cryptography based on the difficulty of large integer factorization. The SM9 algorithm is more efficient than ECC in resource-constrained environments (e.g., IoT devices) due to its smaller key size. The decentralized management and traceability of blockchain technology can meet the demand for tamper-proof property management systems. Blockchain technology, especially the technology used in cryptocurrencies such as Bitcoin, requires a significant amount of computing power to solve complex mathematical problems, verify transactions, and create new blocks. Each node in the blockchain network needs to store a copy of the entire transaction history, which requires a large amount of storage space as the blockchain grows. Therefore, this study proposes a novel SM9-Blockchain identity confirmation system that integrates the state secret SM9 algorithm with Distributed Ledger Technology (DLT) from blockchain.

Research question: This study aims to address the security vulnerabilities, vulnerabilities to attacks, and forgery in existing user identity verification methods for property management systems, and explore a more secure, efficient, and scalable authentication method.

Research objective: Propose a dual-encryption user authentication model that combines the state-secret SM9 algorithm with blockchain technology to enhance authentication security and privacy. Design and implement a user identity verification system based on the SM9 algorithm and blockchain technology to reduce system latency and improve authentication accuracy and efficiency.

Research significance: This study proposes a new dual-encryption authentication method by combining the state-secret SM9 algorithm with blockchain technology, providing a more secure solution for property management systems. This model not only effectively prevents identity forgery and information tampering, but also enhances system reliability and security through the decentralized and immutable nature of blockchain. Furthermore, this study explores how to optimize computational efficiency and communication latency during the authentication process in resource-constrained

*Corresponding author.

environments, which has important theoretical and practical implications.

The research's innovation is reflected in three key aspects. First, a dual-encryption identity authentication model was constructed, combining the state secret SM9 algorithm with blockchain DLT. This overcomes the limitations of existing methods that rely solely on a single encryption or storage mechanism. Second, a time-decay-based trust management mechanism for Certificate Authorities (CAs) was proposed. This mechanism dynamically adjusts the trustworthiness of CAs based on historical behavior, effectively suppressing the repeated issuance of certificates by low-reputation CAs and improving the reliability of the authentication system from the source. This design breaks away from the static model of traditional trust management methods, making the system more resilient to low-reputation CAs. Finally, the research introduced multi-layer blockchain data compression and snapshot technology, enabling efficient collaboration between the main chain and auxiliary chain/off-chain data.

The research's contribution lies in proposing a dual-encryption identity authentication framework that combines the SM9 identity-based encryption algorithm with blockchain-based DLT, overcoming the limitations of traditional single-layer encryption or storage mechanisms. In experiments, the SM9-blockchain model achieved a response latency of 6ms against SQL injection attacks, approximately 50% faster than traditional approaches, enhancing security and authentication efficiency. Secondly, the research designed a CA trust management mechanism based on time decay to dynamically adjust the CA's credibility and effectively prevent duplicate certificate issuance by low-reputation institutions. In experiments, the mis-issuance rate of certificates from low-reputation CAs was reduced from the traditional 20% to 12%, improving the reliability of the authentication system. Finally, a multi-layer blockchain data structure supporting compressed snapshots was constructed, improving data storage and processing efficiency in resource-constrained IoT scenarios. When processing 1,000 concurrent requests, the SM9-blockchain model achieved a response time of 11 ms, approximately 50% faster than traditional blockchain systems, while also reducing energy consumption by 30%.

The structure of the study is divided into six sections. Section II is a literature review. Section III is the methodology, which designs a user identity confirmation model based on DLT and state secret SM9 algorithms. Section IV is the performance validation, which analyzes the performance of the proposed model through experiments. Section V is a discussion, which analyzes the research results. Section VI is the conclusion, which summarizes the research results and shortcomings.

II. Related Work

As the Internet grows, there is an increase in the focus on user data security and privacy across fields like business registration and corporate governance [6]. Blockchain, a leading encryption and traceability tech, is under intensive study by researchers [7]. Attaran's research underscores blockchain's secure encryption, suggesting its use in crypto transactions. It notes that blockchain nearly prevents data alteration or removal without others' private keys, marking a key trend in future crypto security [8]. Li et al. suggested a blockchain framework for encrypting info exchanges among IoT devices, leveraging blockchain's consensus mechanism, encryption, and smart contracts to secure interactions between smart devices [9]. Panda et al. created a blockchain-IoT architecture that secures data privacy and IoT security by using hash chains for key management, eliminating the need for third-party key handlers. The scheme includes an efficient key generation method for mutual authentication through self-verification [10]. To address the limitation of lack of flexibility and responsiveness to dynamic changes of static smart contracts in blockchain applications, Saputra et al. based on the innovative concept of dynamic smart contracts in key-value format framework, AniraBlock, which enhances the data management and supply chain transparency, and achieves the optimisation of data integrity and operational efficiency in the agriculture sector [11]. Saputra et al. further proposed a blockchain-based key-value store that explores interoperability, dynamic data processing and performance testing of smart contracts, thereby enabling dynamic smart contract interactions in the agriculture sector [12].

State secret algorithms are cryptographic algorithms designed and developed independently by China, currently an open source including SM2 to SM9. There are a large number and wide range of SM9-based studies on the Internet. For example, Yang et al. proposed a novel identity-based blind signature scheme combined with message recovery technology to reduce the transmission bandwidth of the signature information by hiding the message in the signature in response to the problem of high resource consumption of the SM9 blind signature scheme in practical applications [13]. Lai et al. introduced an improved online/offline signature scheme for SM9 to accelerate its signing process. The new scheme boosts signature generation speed by 9.9%, reducing signature time to under 1 millisecond, and demonstrates superior performance over existing fast signature algorithms [14]. Aiming at the problem that civil navigation information of BeiDou satellite navigation system (BDS) is vulnerable to spoofing and replay attacks due to the lack of authentication mechanism, Wu et al. proposed a secure authentication protocol BDSec based on the design of China's SM series of algorithms [15]. Zhang and Ye developed a privacy-preserving method leveraging SM9 and blockchain. It employs conditional anonymous ring signatures for enhanced security and optimizes key negotiation using the discrete logarithm assumption, making it faster. The scheme proved to be effective and practical in experiments [16].

In summary, SM9 algorithm and blockchain technology are current research hotspots, and the combination of the two has great potential in protecting blockchain communication and data privacy. SM9 algorithm, as an identity-based cryptographic algorithm, provides a secure key negotiation and authentication mechanism. However, existing research has largely focused on using either the SM9 algorithm or blockchain technology alone, while the combination of the two has not been fully explored. The scalability and efficiency issues of combining the SM9 algorithm and blockchain technology in practical applications remain unaddressed. Furthermore, many studies assume an ideal network

environment, overlooking performance bottlenecks in resource-constrained and high-concurrency environments. Therefore, this research aims to combine the SM9 algorithm with blockchain technology to build a more efficient, secure, and scalable identity verification system, addressing the challenges faced by existing approaches in practical applications.

III. Identity Confirmation Property Management System Based on State Secret SM9 Algorithm and DLT

The study introduces a user identity verification model on DLT to ensure data integrity and accuracy through a decentralized network. It also incorporates the SM9 algorithm for double encryption of the identity verification model and uses asymmetric encryption to safeguard user identity information.

A. Construction of Identity Confirmation Property Management Model Based on DLT

DLT technology is a data storage and recording technology, which allows multiple participants to jointly maintain a growing list of data records. DLT ensures the data consistency of all participants through consensus mechanism. Blockchain DLT stores data in block form, with each block holding transactions and linked via hashing to the previous block, creating a chain structure [17-18]. Once data is recorded, all nodes synchronize it, making alterations to a single block ineffective [19]. DLT also features encryption for secure and confidential data storage. In this, the expression of DLT block is shown in Eq. (1):

$$block = h(data, timesiamp, previous_hash, hash) \quad (1)$$

In Eq. (1), $data$ denotes data stored on the blockchain, which can be interpreted as user identity information or transaction records, etc. $timesiamp$ represents a timestamp used to maintain the order of blocks. $previous_hash$ represent the hash value of the previous block; $hash$ represents the hash value of the current block; $h(.)$ represents the connection function between different blocks, which is used to link blocks and ensure that each block is closely connected to the previous block, forming a chain. Eq. (1) describes the structure and linkage of individual data blocks in a blockchain. In practice, each block is like a "ledger page", in which the user's identity data and operation records are recorded. By linking the hash value to the previous block, the data can be prevented from being tampered with in the middle of the process, thus guaranteeing the security and non-repudiation of the system. The blockchain architecture in this study is constructed based on the Hyperledger Fabric platform, adopting the coalition chain model, through the improved Byzantine Fault Tolerance (BFT) algorithm as the consensus mechanism, combined with the main chain-side chain structure and compressed snapshot technology, which improves the processing performance and data synchronization efficiency while ensuring that the transactions are not tampered with. Each node constructs the block header and synchronously updates the blockchain through the hash function, forming a complete DLT, according to a special combination, as shown in Fig. 1.

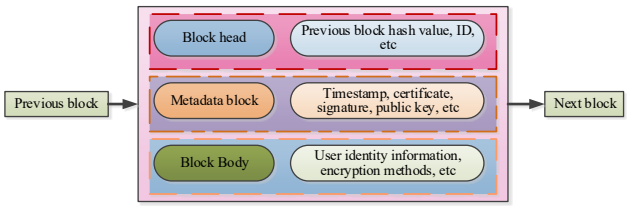


Fig. 1. Schematic diagram of blockchain combination structure.

From Fig. 1, it can be seen that in cross-system identification, the length of the authentication path and the size of the integer volume are the key factors affecting the difficulty of constructing the trust path. In order to adapt to the higher requirements of energy consumption and storage efficiency in large-scale application scenarios, the optimized DLT technology adopted by the institute introduces a lightweight consensus mechanism and compressed storage structure in its design. In terms of energy consumption control, the system adopts an improved BFT algorithm to replace the traditional proof-of-work mechanism to avoid a large number of invalid calculations. Compared with the traditional BFT algorithm, this study simplifies the node election and message transmission mechanisms by introducing pre-selected nodes, thereby reducing redundant broadcast communication and lowering network overhead while maintaining fault tolerance. Compared with commonly used consensus mechanisms such as Proof of Work (PoW) and Proof of Stake (PoS), BFT offers higher determinism and efficiency in consortium blockchain environments. Although PoW provides strong security, it relies heavily on computational resources and consumes a large amount of energy, making it unsuitable for identity authentication systems that require real-time response and resource efficiency. PoS, while improving performance to some extent, depends on economic assumptions in its incentive and security models, which are less compatible with uniformly managed system architectures. In contrast, BFT does not rely on computational competition and offers advantages such as low latency, high controllability, and strong determinism, making it more suitable for the trusted identity authentication model proposed in this study.

In terms of data storage, by setting the validity period and update cycle of identity data, and combining layered storage and snapshot mechanism, only key identity information and index data are retained on the main chain, while the rest of the auxiliary information is distributed to the side chain or off-chain storage. A snapshot mechanism is one that periodically records a full copy of the current state of the blockchain system and transfers the historical block data to off-chain or auxiliary chain storage. To optimize storage requirements, the system uses a snapshot mechanism. By periodically generating snapshots of the blockchain data, the system saves the current state and the historical data is transferred to off-chain storage or external storage. This strategy not only reduces the storage pressure on the blockchain main chain, but also ensures the consistency and integrity of the data through hash values and timestamps. While preserving the current valid data, the historical data can still be traced back through the verification

mechanism to ensure the security and traceability of the data, thus effectively reducing the storage requirements and improving the query efficiency.

In this way, while ensuring data integrity and traceability, it effectively reduces the local storage pressure and bandwidth demand of the nodes, and improves the scalability and green computing capability of the system. Among them, CA auditing is an important work in blockchain. If CA CA_1 has issued M certificates in a period of time, among which there are m legitimate certificates. The weight attenuation of CA certificates, the expression of the legitimacy probability of institutional certificates, and the calculation of credibility are shown in Eq. (2):

$$\begin{cases} f_{PRE}(CA) = \gamma^{t_n - t_{pre}} \\ P_{le} = \dfrac{M}{m}\gamma^{\frac{1}{M}} \\ P_{re} = \dfrac{1}{n}\sum_{i=1}^{n}\left[\alpha f_{PRE}(CA) t_{pre} + \beta P_{le}\right] \end{cases} \tag{2}$$

In Eq. (2), $f_{PRE}(CA)$ represents the CA's weight decay function, reflecting the gradual decrease in trust over time. γ indicates the decay factor; t_n indicates the current time; t_{pre} indicates the moment when the CA issues the last certificate. α, β represents the weight parameter, which has the relationship of $\alpha + \beta = 1$, and both parameters are constants greater than 0 and less than 1; n represents the number of times the organization has been audited. Eq. (2) is used to calculate the credibility of the CA. This equation is based on a time decay model. If the CA issues low-quality certificates or issues certificates frequently, the CA's credibility gradually declines over time. The study introduces a 'trust decay' mechanism for dynamically assessing the trustworthiness of CA organizations. The system records the frequency, legitimacy rate and historical behavior of certificate issuance by CAs, and sets a time decay factor to make the trust value of CAs decay naturally over time, so as to prevent the risk of abuse brought by long-term accumulation of trust. If a CA frequently issues illegal or low-quality certificates within a certain period, its weight will be rapidly reduced, and it will eventually be regarded by the system as a low-trust or failing organization. This mechanism improves the resilience and robustness of the certification system and effectively prevents the risk of a single point of failure of the authority center. The specific calculation equation is shown in Eq. (3):

$$\begin{cases} S = HA(s) \\ r = (S + x) \bmod n \\ k = (1+v)^{-1}(t - rt) \bmod n \end{cases} \tag{3}$$

In Eq. (3), S denotes the encrypted ciphertext; $HA(.)$ denotes the hash encryption function; s denotes the plaintext to be encrypted. x denotes the parameter of the public key; $\bmod n$ denotes that the mode of this calculation is elliptic curve. k denotes another parameter of elliptic curve digital signature, the larger the value, the higher the strength of the signature, and the higher the security. v denotes the parameter in the private key; t denotes the moment when this calculation is performed. Eq. (3) is the calculation for cryptographic signing of user identity information, which indicates that the user encrypts the information to be signed using a private key to generate a signature. The signature can be used to verify the legitimacy of the user's identity in subsequent steps. In practice, the user's identity information is converted into an encrypted digest, and the equation generates the signature parameter r. Once the server receives the signature, it can verify the legitimacy of the identity and avoid identity forgery. The proposed model uses digital signatures to verify user identities on the Public Key Infrastructure (PKI). PKI employs asymmetric encryption with a public-private key pair for secure data transmission, where the public key is shared for encryption and the private key is kept secret for decryption. Automated certificate lifecycle management in PKI systems reduces complexity and errors in manual operations. When a user from PKI domain 1 wants to authenticate with a server in domain 2, they must submit a certification request, including a digital signature and blockchain certificate to the server in domain 2. The server then verifies the user's identity with the certificate issuer [20]. Fig. 2 illustrates the steps of user authentication.

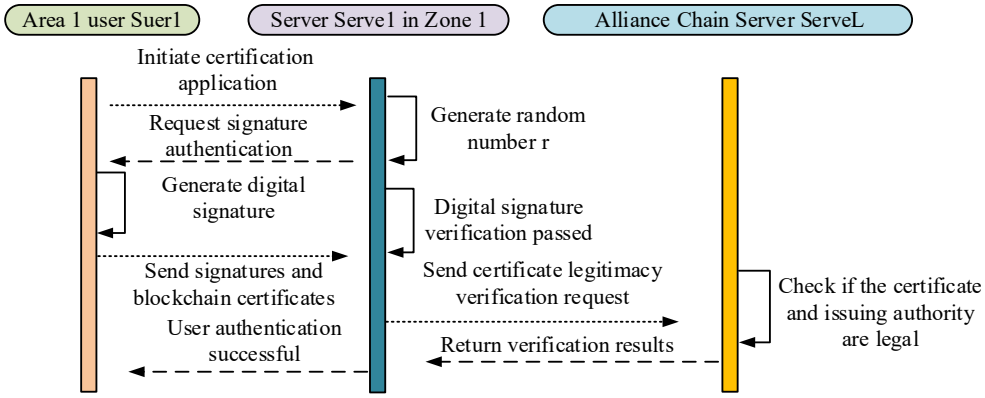


Fig. 2. Schematic diagram of user identity authentication process.

Fig. 2 illustrates the PKI encryption process: the sender gets the receiver's public key, encrypts data with it to create ciphertext, and sends the ciphertext. The decryption process involves the receiver using their private key to decrypt the ciphertext and retrieve the original data. Leveraging blockchain's decentralized and tamper-proof features, the study develops a secure user identification model offering efficient authentication, authorization, and protection of user privacy and security. The interaction mechanism between the blockchain and the SM9 algorithm is shown in Fig. 3.

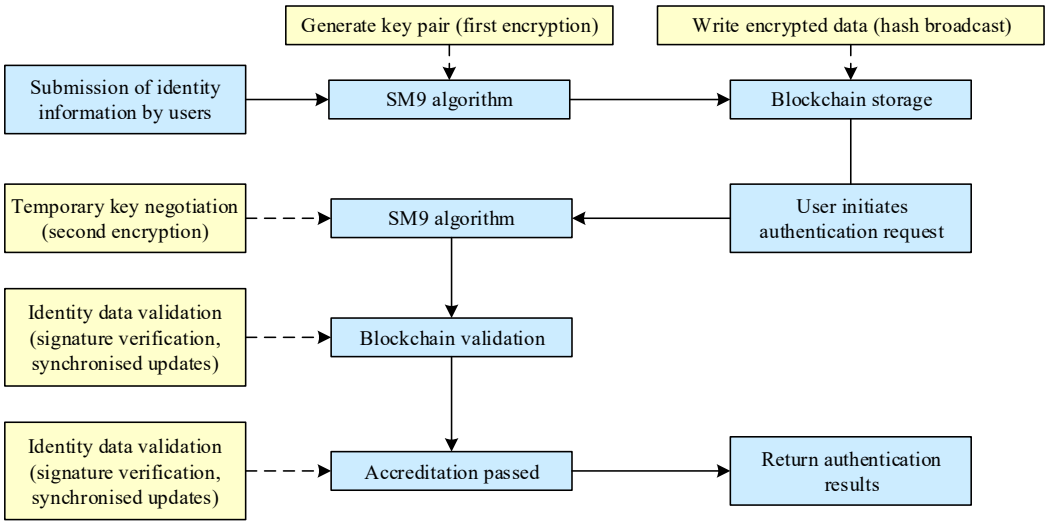


Fig. 3. Technical interaction mechanism between blockchain and the SM9 algorithm.

Fig. 3 shows how the SM9 algorithm generates key pairs and encrypts the data, how the blockchain stores the encrypted data and verifies it, and ultimately ensures the security and efficiency of the system through key negotiation and authentication. And in order to ensure the security of the off-chain data storage, the Institute lifting system has adopted a variety of technical measures. All off-chain data is encrypted before storage, and the SM9 algorithm is used for secondary encryption. At the same time, distributed storage and data slicing technologies are used to increase data redundancy and reliability and prevent a single point of failure. Managing permissions through role-based access control ensures that data access is restricted to authorized users only. In addition, the system introduces audit logs to record all data operations to ensure data traceability.

B. Secondary Encryption of Identity Confirmation Physical Model Based on SM9 Algorithm

In order to improve the encryption performance of the identity verification property management model, the use of the national security SM9 algorithm for secondary encryption is studied. The SM9 algorithm typically uses an asymmetric encryption system, also known as a public key cryptosystem, to generate identity-based public key encryption schemes to generate key pairs. Among them, the expressions for calculating the public and private keys are shown in Eq. (4):

$$\begin{cases} Pu_i = Hash(id_i) \\ Pr_i = \left(\dfrac{w}{w + Pu_i} \right) Pu_i \end{cases} \tag{4}$$

In Eq. (4), Pu_i denotes the public key of the first i user; $Hash(.)$ denotes the hash encryption function; id_i denotes the identity information of the first i user. w represents the master private key generated by the public key generator; Pr_i represents the private key of the i user. Eq. (4) is the basis for the SM9 algorithm to generate user public-private key pairs. The system generates its own encryption key based on the user's identity information, which is equivalent to "using identity information directly into a key". During key pair generation for user authentication, the IMS generates a random large prime number. This prime is used to create a temporary public and private key, as detailed in Eq. (5):

$$\begin{cases} \phi = N_1 Pu_i \\ \theta = N_2 \Pr_i \end{cases} \tag{5}$$

In Eq. (5), N_1 denotes the randomly generated large prime number; ϕ denotes the temporary public key. N_2 denotes the randomly generated large prime number, which is not necessarily the same as the value of N_1 ; θ denotes the temporary private key in the key negotiation process. During the key negotiation process, both parties generate temporary key pairs, which are used only for encryption tasks in the current session. The use of temporary keys reduces the security risks associated with using the same key for a long period of time. After generating the temporary key, the initiator sends the encrypted ciphertext and the temporary public key to the authenticator. The authenticator decrypts the message and checks the bilinear pairing conditions. If conditions are met, they calculate the session key, as detailed in Eq. (6).

$$\begin{cases} E(\theta, s' + Pu_i N_1) = E(\phi, s') \\ sk = Hash(id_i \| \phi \| \phi' \| \beta) \\ \partial = N_1 \phi \end{cases} \tag{6}$$

In Eq. (6), s' denotes the master public key generated by the public key generator. ϕ' denotes the temporary public key of the authentication server; ∂ denotes the private key base point in the system parameters. Generating the session key is the final step in the key negotiation phase, and the detailed flow of key negotiation is shown in Fig. 4.

Fig. 4 shows the detailed steps of the key negotiation process based on the SM9 algorithm. Firstly, both parties generate their own temporary public-private key pairs for initialization. Secondly, both parties exchange parameters for key negotiation in this phase, such as algorithm version or key size, to ensure that they are communicating under the same standard. Subsequently, in the 'Temporary Key Negotiation' phase, both parties generate a temporary key based on the exchanged parameters, which is only used for the current session to enhance the security of the communication. Verify that the temporary key is correctly authenticated by both parties. If the authentication fails, the process returns to the "Parameter Exchange" phase and restarts the entire process to prevent man-in-the-middle attacks and to ensure that the key is secure before it is used. If authentication passes, the process continues to the Key Negotiation Calculation phase, where both parties calculate the final shared key based on the temporary key and possibly other parameters, ensuring that the key is consistent between the parties and can be used for encrypted communication. In the 'Encrypted Communication' phase, both parties use the negotiated key for encrypted communication to ensure that the data is secure from unauthorised access during transmission. Subsequently, in the 'key validation' phase, the parties confirm that the key has been used correctly to ensure that there are no errors or tampering. At this point, the blockchain user identity confirmation property management model SM9-Blockchain optimized based on the SM9 algorithm runs, as shown in Fig. 5.

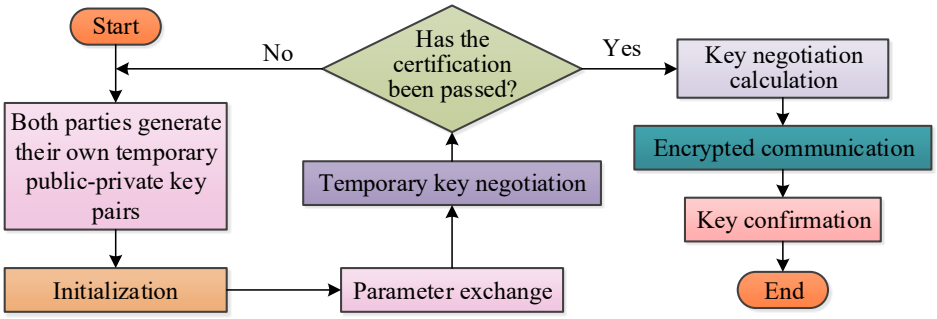


Fig. 4. Detailed process of key negotiation.

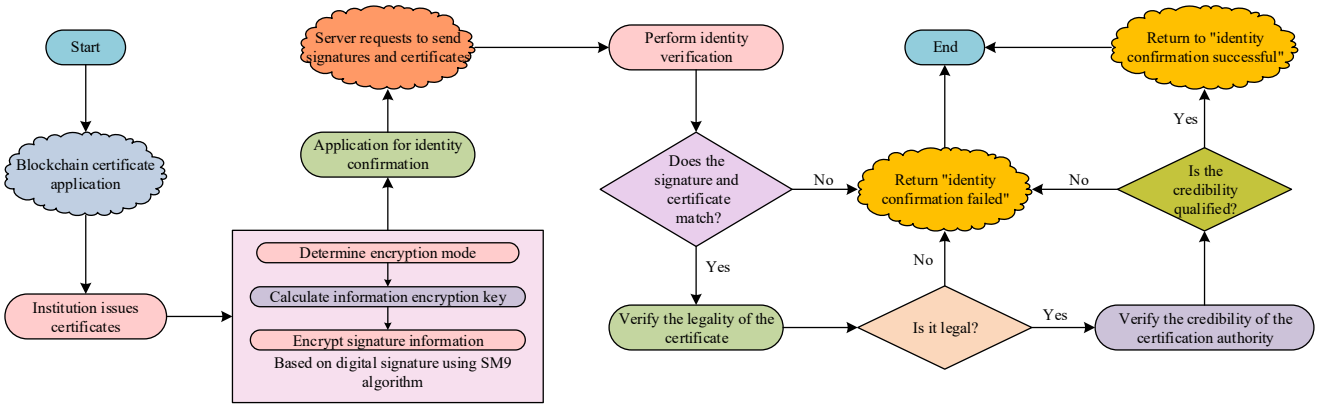


Fig. 5. SM9 blockchain model running process.

Fig. 5 shows the running process of the optimized SM9-blockchain model. First, the two parties exchange parameters used for key negotiation, and the user submits an authentication request through the blockchain to complete the authentication. At the same time, the system uses the SM9 algorithm to double-encrypt and verify the user's identity. After the authentication is passed, the user data is securely stored on the blockchain. Next, key negotiation is performed between the user and the system to generate a session key. Encrypted communication using the session key ensures the security of data transmission. This model makes use of the tamper resistance of blockchain and the encryption strength of SM9 algorithm to improve the security and efficiency of user authentication.

IV. SM-Blockchain Model Performance Validation

A. SM9-Blockchain Model Performance Validation

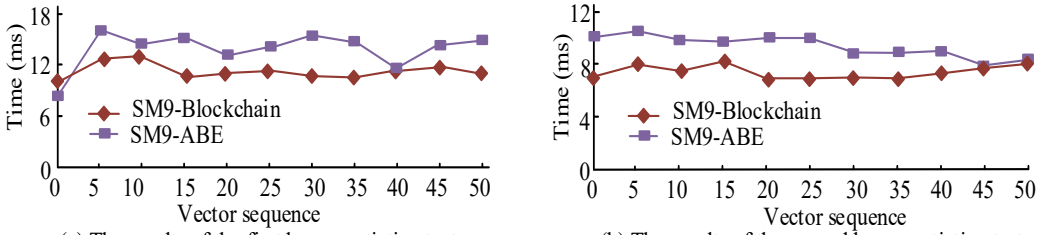
To validate the performance of the proposed model, the study conducted experiments in the hardware and software environments shown in Table I.

TABLE I. Configuration Table of Software/Hardware Experimental Environment

Item	Configuration	Item	Configuration
Operating system	Ubuntu 16.04	Programming language	Java
Processor	8-core CPU	Key generation tools	Cryptogen & Fabric-CA
Memory	16 GB RAM	Key generation method	Combination of static and dynamic
Blockchain platform	Hyperledger Fabric 1.4	Chaincode deployment	Automated via network.sh script

In this case, the decay factor γ is 0.1 and the weight parameter α, β is 0.5 [21-22]. The SM9 algorithm, as an identity-based cryptosystem, is faster in key generation and encryption and decryption than traditional public key systems. Therefore, the study chooses test values that can reflect the efficiency of the SM9 algorithm. The study follows the three principles of reproducibility, practicability, and comparability in the selection of test values. First, in order to ensure that the test data can truly reflect the performance of the algorithm, the parameters selected in the test (elliptic curve order, hash function bits, temporary key length, etc.) are based on the national commercial cryptography standard and the recommended configuration of Hyperledger Fabric. Second, when simulating the actual operating environment of the blockchain, a typical data load size is set to simulate the transmission performance under medium load in reality. Again, to enhance the interpretability of the results, the selected test vectors include Padding, Differential Analysis, and Random, and compare the commonly used model SM9-ABE. Parameter values that can simulate the actual blockchain operating environment were selected in the tests to evaluate the performance and energy consumption of the model in real applications. Parameter values that can ensure transaction security while minimizing the impact on system performance are selected.

The study evaluates the SM-Blockchain model's performance using three test vector sets: Padding Test Vectors (PTV) for testing encryption and decryption correctness and efficiency, Differential Cryptanalysis Test Vectors (DCTV) for assessing resistance to attacks, and Random Test Vectors (RTV) for performance and security. The above vectors are all derived from the publicly available national cryptography SM9 algorithm standard test set and open source implementation, mainly including the National Cryptography Administration SM9 standard and test vectors (http://www.gmbz.org.cn/main/viewfile/20180108015305901286.html) and the GmSSL open source national cryptography algorithm library (https://github.com/guanzhi/GmSSL). The SM9-ABE framework is included for comparison. Fig. 6(a) shows that the establishment of a communication connection or initiating a validation process speeds for SM9-ABE and SM9-Blockchain on various input vectors in the first test. SM9-Blockchain averages at 11.07 ms, while SM9-ABE at 14.11 ms. Fig. 6(b) reveals average times of 7.96 ms for SM9-Blockchain and 9.88 ms for SM9-ABE. This result shows that the SM9-Blockchain model is superior to the SM9-ABE model in the speed of establishing a communication connection or initiating a validation process. In addition, the research simulates different network conditions and data loads through automated scripts, which ensures the accuracy and reliability of test results.

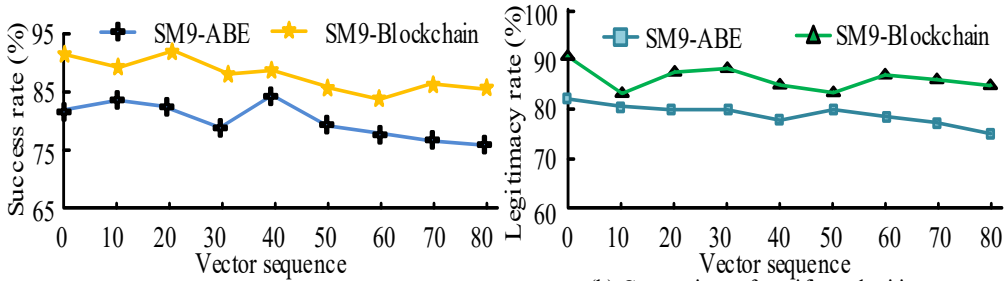


(a) The results of the first key negotiation test (b) The results of the second key negotiation test

Fig. 6. Comparison of models establishes a communication connection or initiates a validation process speed.

Fig. 7 illustrates the success and legitimacy rates for key negotiations in the SM9-ABE and SM9-Blockchain models under the PTV vector set. As can be seen from Fig. 7(a), the success rate of key negotiation of SM9-Blockchain model is 88.73%, which indicates that this model can effectively establish secure session keys between users in most cases. In contrast, the success rate of SM9-ABE model is 83.77%, which is slightly lower than that of SM9-Blockchain model. This may be because the SM9-Blockchain model adopts more optimized algorithms and protocols in the process of key exchange. The difference of success rate may be due to the fact that SM9-Blockchain model provides a more stable network environment and a more reliable data synchronization mechanism by using the distributed characteristics of blockchain technology, thus improving the stability and success rate of key agreement. In Fig. 7(b), the certificate validity rate of the SM9-Blockchain model is 87.51%, which means that the model can verify the validity and legality of the certificate with high accuracy. The certificate validity rate of SM9-ABE model is low, which may be due to some loopholes or deficiencies in the process of certificate management and verification, which leads to the decrease of the accuracy of validity verification. The high

certificate legitimacy rate of SM9-Blockchain model benefits from the non-tampering and transparency of the blockchain, and the issuance and status of each certificate can be tracked and verified on the blockchain, thus effectively preventing forgery and tampering.

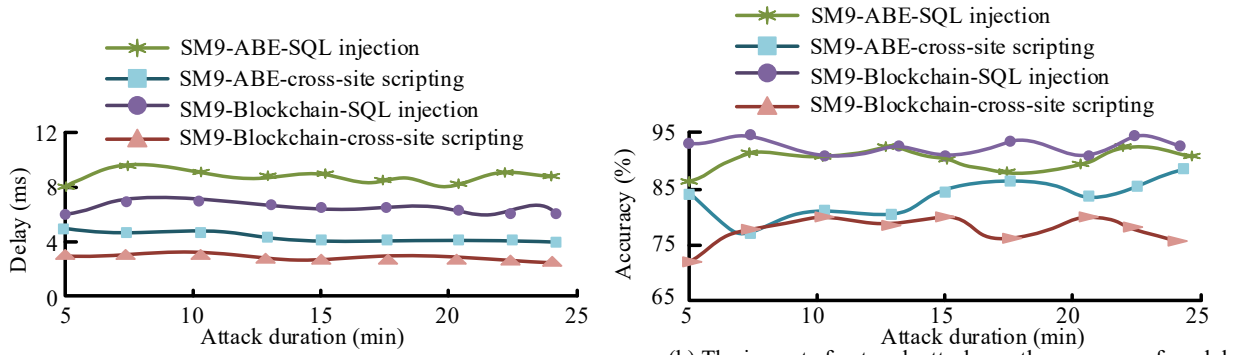


(a) Comparison of key negotiation success rates (b) Comparison of certificate legitimacy

Fig. 7. Comparison between the success rate of model key negotiation and the legality rate of certificates.

B. SM9-Blockchain Model Security Validation

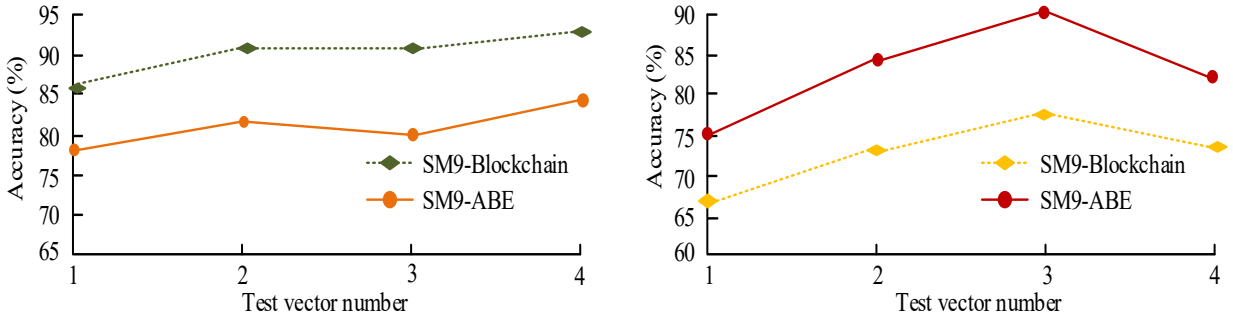
The study tests the model's security against SQL Injection and XSS attacks. Fig. 8(a) shows the SM9-Blockchain model experiences about a 6ms delay under SQL Injection and a 3ms delay under XSS attacks. Fig. 8(b) indicates that the SM9-Blockchain model has a higher average correct rate than the control model under various attacks.



(a) The impact of network attacks on model runtime latency (b) The impact of network attacks on the accuracy of model identity confirmation

Fig. 8. Model security verification.

To analyze the factors affecting user identity confirmation accuracy, the model uses PTV and RTV vectors. Fig. 9(a) shows that SM9-ABE has an average correct rate of 81.29%, while SM9-Blockchain has 90.41%. Fig. 9(b) reveals that SM9-ABE's average correctness is 71.81% and SM9-Blockchain's is 83.25%.



(a) The recognition accuracy of random vectors generated in RTV (b) The recognition accuracy of random vectors generated in PTV

Fig. 9. Comparison of user identity confirmation accuracy.

To analyze encryption and decryption delays in user identity confirmation under normal scenarios, experiments were conducted using the PTV vector set. In Fig. 10(a), the encryption delay of SM9-Blockchain model is about 0.35 ms lower than that of SM9-ABE model under the same conditions. This improvement may be attributed to the adoption of more efficient algorithm and optimized data structure in the encryption process of SM9-Blockchain model. The lower encryption delay means that the SM9-Blockchain model is faster in dealing with encryption tasks, which is especially

important for application scenarios that need rapid response, such as real-time communication and online transactions. The SM9 algorithm, the core cryptography of the model, is important for reducing latency due to its smaller key size and faster encryption/decryption speed. Lower cryptographic latency means that the SM9-Blockchain model is faster in processing cryptographic tasks, which is especially important for application scenarios that require fast responses (e.g., real-time communication and online transactions). The distributed architecture of blockchain technology, however, improves the efficiency and reliability of data synchronization, which reduces delays in the user identification process. Each node stores a copy of the blockchain, and this decentralized structure reduces the single point of failure and bottlenecks that can occur in centralized systems. It can also be seen from Fig. 10(b) that the delay of the proposed model is about 0.39ms lower than that of the SM9-ABE model. This shows that the SM9-Blockchain model not only performs well in the encryption stage, but also provides fast processing ability in the decryption stage. Decryption delay is very important to ensure real-time data and improve user experience, especially in the environment of frequent data exchange. In this study, several key security protocols are introduced in the contract design. First, when a user submits an authentication request, the system uses the signature verification function built into the on-chain contract to verify the request information to ensure that the request source is real and not forged. Second, by introducing the anti-replay mechanism based on timestamp and Nonce, attackers are effectively prevented from utilizing historical requests to initiate pseudo-authentication operations. In addition, the smart contract restricts the invocation scope of contract functions through role authority control to prevent illegal access to sensitive resources. To address the risk of contract logic vulnerabilities, Mythril, a static code analysis tool, is introduced to conduct vulnerability scanning before the system is deployed, and combined with the test network for simulated attack verification, to ensure that the contract has a strong anti-re-entry and anti-overflow capabilities. The above mechanisms together constitute the basic security protocol framework used in this system to defend against typical attacks, which improves the robustness and credibility of the system in real application scenarios.

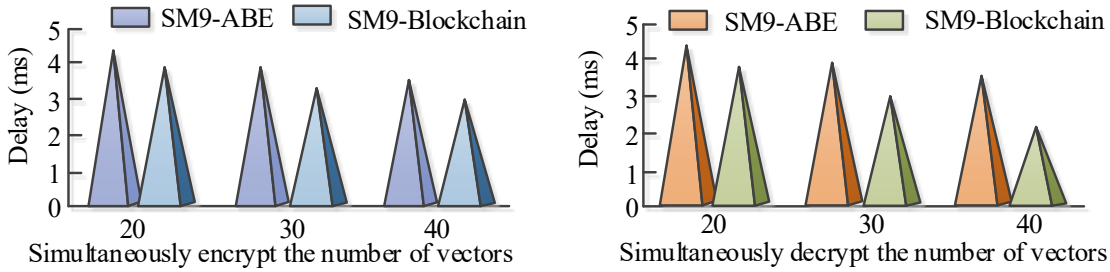


Fig. 10. Comparison of encryption and decryption latency among different models.

C. Performance Comparison of Different Methods

The study compares blockchain-based authentication management methods proposed by other scholars, including Miao et al.'s [23] blockchain-enabled privacy-preserving authentication management protocol for the Internet of Medical Things, Kumar et al.'s [24] blockchain-based authentication method, and Kang et al.'s [25] blockchain-based authentication scheme for an enhanced and lightweight medical sensor network. It benchmarks the time overhead for key generation, proving keys, and verifying keys in the PTV set for these methods. Details are shown in Fig. 11.

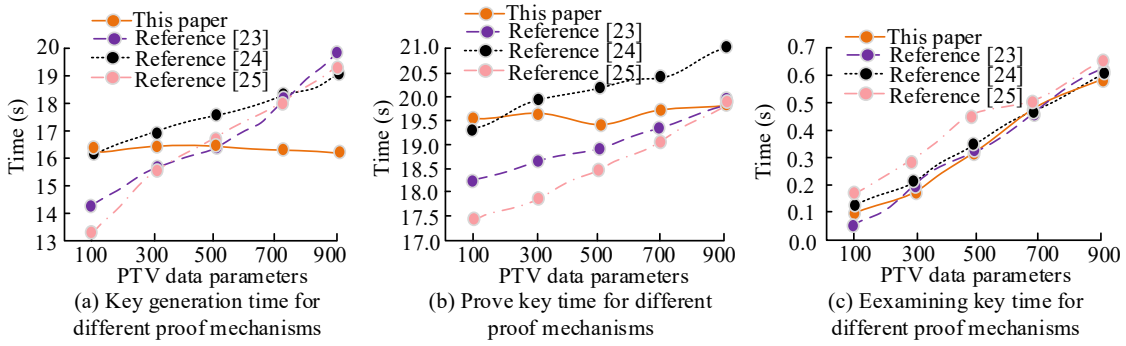


Fig. 11. Comparison of the time overhead required for key generation and verification under different authentication mechanisms.

Fig. 11(a) shows that SM9-Blockchain's key generation time is consistently around 16.29s, with an average reduction of 3.93%, 7.39%, and 1.51% compared to other methods. Fig. 11(b) indicates that SM9-Blockchain is least affected by data parameter size in key proving, while other methods show an increasing trend with larger parameter sizes, possibly due to larger proof key storage requirements. Fig. 11(c) shows that SM9-Blockchain's verification time is reduced by 4.84% on average when handling 900 PTV data parameters. These results demonstrate that SM9-Blockchain outperforms others in encryption and decryption time. The performance comparison results of key indicators of the four methods are shown in

Table II. To ensure the statistical reliability of the results, each test was conducted no fewer than twenty times under the same conditions, and all metrics are reported as mean ± standard deviation. Performance differences were evaluated for statistical significance using paired t-tests with a significance level of $\alpha = 0.05$.

TABLE II. PERFORMANCE COMPARISON RESULTS OF KEY INDICATORS OF FOUR METHODS

Index	SM9-Blockchain	Miao et al. [23]	Kumar et al. [24]	Kang et al. [25]
Average establish a communication connection or initiate a validation process time (ms)	11.07±0.22	14.52±0.48	12.31±0.35	13.83±0.41
Average key negotiation success rate (%)	88.73±1.15	85.20±1.60	87.11±1.42	86.54±1.51
User identity confirmation accuracy (%)	90.41±1.07	87.61±1.32	88.93±1.21	89.26±1.29
Delay time under SQL injection attack (ms)	6.00±0.31	8.22±0.37	7.52±0.29	7.81±0.33
Delay time under XSS attack (ms)	3.00±0.18	4.53±0.22	3.81±0.20	4.04±0.21
Average key generation time overhead (s)	16.29±0.47	18.51±0.55	17.10±0.52	17.94±0.54
Average key proof time overhead (s)	19.52±0.62	21.33±0.72	20.25±0.68	20.82±0.70
Average key verification time overhead (s)	0.36±0.05	0.45±0.06	0.40±0.05	0.42±0.06

Through the comparison in Table II, it can be seen that the SM9-Blockchain model performs better than other models in most key performance indicators. Especially in the accuracy of user identity confirmation and the delay time under SQL injection and XSS attacks, the SM9-Blockchain model shows obvious advantages. This shows that the SM9-Blockchain model has obvious advantages in improving security and efficiency. To further evaluate the performance advantages of the SM9 algorithm in identity authentication systems, this study conducts a comparative analysis with traditional encryption protocols, including Rivest–Shamir–Adleman (RSA) (1024-bit), Elliptic Curve Cryptography (ECC) (256-bit), and Advanced Encryption Standard (AES) (128-bit). All tests were conducted under the same environment and transaction volume. This is shown in Table III.

TABLE III. PERFORMANCE COMPARISON BETWEEN SM9 AND CONVENTIONAL ENCRYPTION ALGORITHMS

Algorithm	Key length	Key generation time (ms)	Encryption latency (ms)	Decryption latency (ms)	CPU utilisation (%)	Memory usage (MB)	Energy consumption (J)
SM9	256 bit	16.29	7.81	8.70	12.34	25.43	0.09
RSA	1024 bit	31.20	9.56	18.40	20.15	38.62	0.18
ECC	256 bit	23.45	8.63	12.51	17.98	33.24	0.12
AES	128 bit	1.52	4.13	4.05	8.56	14.79	0.03

As can be seen from Table III, RSA exhibits significantly higher average latency in key generation and decryption compared to SM9 (16.29 ms and 8.7 ms). ECC performs better than RSA but remains slightly slower than SM9 under concurrent transaction loads due to computational bottlenecks in key negotiation. AES, as a symmetric encryption algorithm, delivers excellent speed but lacks support for identity binding and key negotiation, making it unsuitable for standalone use in blockchain-based identity authentication. Although the symmetric encryption algorithm of AES has the lowest CPU usage (8.56%), considering the strong encryption strength provided by the SM9, the SM9's CPU usage is still within a reasonable range, which is a superior price/performance ratio, especially with its high level of security. Overall, SM9 strikes a balance between security and efficiency by enabling identity-based encryption, digital signature, and key exchange with smaller key sizes and lower latency. This makes it particularly suitable for resource-constrained environments, such as IoT devices and lightweight authentication terminals.

D. Practical Validation

To further evaluate the performance of the proposed authentication model in real-world and IoT scenarios, two experimental environments were designed: a real property management system and an IoT device environment. In the real property management system test, the model was deployed on an existing platform for limited-scale validation. The system included multiple users and multiple device nodes to simulate concurrent authentication requests. During deployment, the front-end authentication module interacted with the back-end RESTful API via the HTTPS protocol. The back-end invoked the SM9 double encryption module and Hyperledger Fabric nodes to perform on-chain authentication and verification. The property management system ran on real servers and client devices, with a local area network latency of approximately 20 ms and a bandwidth of 50 Mbps.

In the IoT device test, a Raspberry Pi 3 (1GB RAM, quad-core processor) was used as the end node, connected to IoT sensors and low-power devices. Communication was securely established with the IoT gateway via an MQTT protocol adapter, with the model also performing encryption, authentication, and verification in the back-end. The IoT devices are connected to the network via Wi-Fi, with a latency of approximately 50 ms and a bandwidth of 10 Mbps.

In both environments, the system first initialized and configured the hardware and network conditions, and then seamlessly integrated with the existing platform (including the PKI certificate system and device management platform) through an interface adaptation layer. The user-side authentication process was triggered with a single click, eliminating the need for manual key management or plugin installation. Interaction latency was kept to under 0.5 seconds,

and the interface style remained consistent with the existing system. Key performance indicators were measured in both environments. Detailed results are shown in Table IV.

TABLE IV. TEST RESULTS OF DIFFERENT ALGORITHMS IN VARIOUS ENVIRONMENTS

Experimental environments	Algorithm	Response time (ms)	Authentication success rate (%)	Network bandwidth consumption (KB/s)
Real-world	SM9	16.97	91.26	22.68
	RSA	28.74	81.32	36.48
	ECC	23.85	86.05	30.92
	AES	18.42	88.14	24.75
IoT	SM9	19.06	87.93	19,47
	RSA	39.21	74.81	30.67
	ECC	33.54	79.49	27.18
	AES	21.13	85.20	21.60

As shown in Table IV, in the real-world property management system test, the response time of the SM9 algorithm was 16.97 ms, which is 40.96%, 28.81%, and 7.88% faster than RSA, ECC, and AES, respectively. This indicates that SM9 has a clear advantage in processing efficiency within practical systems. The authentication success rate of SM9 in this environment reached as high as 91.26%. Combined with the results from the IoT device environment, the superiority of the proposed method is further demonstrated. In the resource-constrained IoT environment, SM9 achieved an authentication success rate of 87.93%, which is 17.52%, 10.62%, and 3.20% higher than RSA, ECC, and AES, respectively. This shows that the proposed method maintains strong authentication performance even on devices with limited resources. Finally, the study further selected two common and representative attack methods to evaluate the security of the system, namely the replay attack and the impersonation attack.

Replay attack refers to a situation where an attacker intercepts and records legitimate authentication requests from users and then resends the same data to the system in an attempt to bypass identity verification. Impersonation attack means that an attacker forges the identity or credentials of a legitimate user or node in order to obtain authentication privileges or access to sensitive resources. The study designed corresponding defense mechanisms based on the SM9-Blockchain model and conducted a comparative evaluation with traditional encryption systems, RSA and ECC. The details are shown in Table V.

TABLE V. SECURITY AND SCALABILITY COMPARISON UNDER REPLAY AND IMPERSONATION ATTACKS

Evaluation Metric	SM9-Blockchain	RSA	ECC
Replay attack detection rate (%)	99.82±0.14	87.20±1.40	91.45±1.25
Impersonation attack detection rate (%)	98.50±0.42	85.70±1.58	89.35±1.31
Authentication throughput (requests/s)	158.42±3.51	112.74±4.15	128.20±3.91
Average authentication latency (ms)	21.85±1.20	36.57±1.82	29.46±1.52
Max concurrent devices supported	1200	750	880

As can be seen from Table V, in the replay attack test, the SM9-Blockchain model achieves a 99.82% interception rate relying on the timestamp and session key mechanism, which has stronger anti-replay capability than RSA and ECC. In the impersonation attack test, SM9 achieves a 98.50% recognition rate based on the identity signature and certificate chain verification mechanism, which is significantly better than RSA and ECC. In addition, in a highly concurrent environment, SM9-Blockchain can still support 1,200 devices to authenticate concurrently, and the authentication throughput reaches 158.42 times/s, with an average latency of only 21.8 ms, which indicates that it has a superior ability to scale. The capability is more superior.

To fully validate the performance of the proposed model under scalability, extreme stress, and long-term operation conditions, the study designed three scenarios: concurrent expansion testing, stress testing, and long-term data accumulation simulation. In the scalability test, the number of concurrent device connections was gradually increased from 200 to 1600. In the stress test, the number of concurrent connections was further increased to 2000 to simulate peak load. In the long-term data accumulation simulation, nodes were continuously run and authentication transactions were continuously written. The storage growth rate and node synchronization latency were recorded after one, three, and six months. The experimental results are shown in Table VI.

TABLE VI. SM9-BLOCKCHAIN MODEL SCALABILITY VERIFICATION

Test scenario	Concurrency level / Duration	Avg. authentication latency (ms)	Authentication throughput (req/s)	Storage growth rate (%)	Node sync delay (ms)
Scalability test	200 devices	19.42 ± 0.38	155.12 ± 3.42	-	-
	800 devices	20.57 ± 0.41	154.03 ± 3.51	-	-
	1200 devices	21.85 ± 0.44	158.42 ± 3.51	-	-
	1600 devices	24.16 ± 0.50	149.35 ± 3.72	-	-
Stress test	2000 devices	27.94 ± 0.55	138.27 ± 3.85	-	-

Long-term data accumulation	1 month	-	-	8.53 ± 0.12	2.83 ± 0.05
	3 months	-	-	24.68 ± 0.21	3.12 ± 0.04
	6 months	-	-	42.37 ± 0.26	3.38 ± 0.06

Table VI shows that in the scalability test, when the number of concurrent devices increased from 200 to 1200, the average authentication latency increased only slightly, while the throughput fluctuated slightly and remained above 150 req/s, demonstrating that the system can operate stably under high concurrency conditions. In the stress test, when the number of concurrent accesses reached 2000, the latency increased to 27.94 ms and the throughput dropped to 138.27 req/s, demonstrating that the system still has sufficient processing capacity under extreme loads. In the long-term data accumulation test, thanks to the multi-layer blockchain and snapshot compression mechanism, the main chain storage growth rate was 42.37% over 6 months, and the node synchronization latency remained in the millisecond range with minimal increase. This demonstrates that the proposed model not only has good concurrency processing capabilities in the short term, but also maintains low resource consumption and stable performance in the long term.

V. Discussion

The SM9-Blockchain model proposed in the study excels in authentication accuracy, encryption and decryption efficiency, and resource adaptability. The constructed double encryption structure and traceable certificate mechanism guarantee the system's lower authentication latency and higher throughput rate in high concurrency and resource-constrained scenarios. In addition, compared with the RSA and ECC models, SM9-Blockchain achieves significant improvements in identity confirmation accuracy and key generation efficiency. From the perspective of mechanism design, the model integrates the structure of the main chain and auxiliary chain, and realizes the collaborative processing of identity authentication and data consistency through the on-chain trusted CA authorization and chain code execution mechanism; the CA trust decay mechanism effectively inhibits the repeated authorization of low-credibility nodes, and the chain code snapshot strategy improves the storage and response efficiency of the system while safeguarding the performance of the main chain. These designs enhance the system's adaptability to complex network environments while improving authentication accuracy. Furthermore, the model's design takes into account internal threats, CA compromise, and practical key management needs. Role-based access control and on-chain auditing are used to prevent internal unauthorized access, trust decay and certificate status broadcasting are used to address CA compromise, and key recovery and revocation mechanisms are provided to ensure long-term security.

In terms of practical application, the system still shows good scalability and platform compatibility. The system can be applied to intelligent communities, IoT terminal access, distributed identity management and other scenarios. Experimental results also demonstrate the model's adaptability and interoperability across diverse regulatory frameworks. Within China's national cryptographic system, SM9 enables authentication compliant with the Cryptography Law, whilst in the US FIPS 140-3 environment, it supports PKI systems including ECC and RSA. Concurrently, the model supports integration with global standards including X.509, OAuth 2.0, and FIDO, while enabling cross-chain interoperability across platforms such as Hyperledger Fabric and Ethereum. Its broad cross-platform and cross-domain applicability has been validated in smart community, IoT, and financial healthcare scenarios. However, with the continuous evolution of technology and the expansion of application scenarios, identity authentication models face new challenges and development requirements in real-world deployments. Firstly, in terms of security, the advancement of quantum computing poses potential threats to existing cryptographic systems. As an identity-based cryptographic scheme based on bilinear pairings, the SM9 algorithm relies on mathematical problems that may no longer be secure under quantum computing. In addition, the hash algorithms and digital signature mechanisms used in blockchain systems are also susceptible to quantum attacks. Therefore, to enhance the model's security adaptability in future quantum environments, it is advisable to introduce post-quantum cryptographic algorithms, such as lattice-based or hash-based signature schemes, in alignment with the standards currently being promoted by institutions such as the National Institute of Standards and Technology of the United States.

Secondly, although this model employs encryption mechanisms and role-based access control policies, there remains a potential risk of privacy leakage during data interaction between on-chain and off-chain environments. Future improvements could incorporate privacy-enhancing technologies such as anonymous credentials and zero-knowledge proofs to mitigate such issues. While blockchain achieves decentralized data storage, PKI still exhibits centralized tendencies during certificate issuance. To address this, the CA trust decay mechanism proposed in this study dynamically adjusts the trustworthiness of certificate authorities based on their historical behavior, thereby effectively reducing the risk of long-term trust monopolization by authoritative institutions. With the development of quantum computing, existing cryptographic algorithms such as SM9 and elliptic curve cryptography may face the threat of being cracked. To enhance the system's security adaptability in future environments, further research into integrating post-quantum cryptographic algorithms should be considered to strengthen the model's long-term robustness in identity authentication scenarios.

Finally, to improve the system's scalability and interoperability, future research will explore cross-chain integration capabilities with other blockchain platforms such as Ethereum and Hyperledger. Through techniques such as relay chains, hash time-locking, oracle services, or pluggable identity modules, the model can enable trusted identity data sharing and verification across multiple blockchains.

VI. Conclusion

This study proposes a dual-encryption authentication framework that combines the SM9 algorithm and blockchain technology. It also designs a time-decay-based CA trust management mechanism and a multi-layer blockchain architecture. Experimental validation demonstrates that the model demonstrates outstanding authentication security, responsiveness, and scalability, achieving a latency of 11.07 ms, a key agreement success rate of 88.73%, and an identity confirmation accuracy of 90.41%. The key contribution of this study lies in its breakthrough of the limitations of single encryption or storage mechanisms by innovatively proposing a dual-encryption authentication framework and optimizing storage and processing efficiency through a multi-layer blockchain architecture. However, the experiments were limited to a simulated environment. Future research could focus on further optimizing the model's performance in resource-constrained IoT environments and enhancing its applicability for large-scale deployments. In summary, this study provides a new solution for authentication in property management systems and IoT devices, improving security, reliability, and scalability, and laying the foundation for further research in related fields.

Acknowledgment

This study was supported by 2022 Nanyang Science and Technology Development Plan Project (NO. KJGG105); Nanyang Medical College 2022 Scientific Research Fund Project (grant No. 2022ZRKX005), Nanyang Science and Technology Plan Project for 2024-2025 (grant No. 24KJGG127).

References

[1] Yusuf M, Yusup M, Pramudya R D, Fauzi A Y, Rizky A, "Enhancing user login efficiency via single sign-on integration in internal quality assurance system (espmi)," International Transactions on Artificial Intelligence, vol. 2, no. 2, pp. 164-172, 2024.

[2] Selvam D, Khanna A, "Enhancing utility sector efficiency and security: Integrating digital identity systems amidst privacy and ransomware challenges," International Journal of Advanced Research in Science, Communication and Technology, vol. 4, no. 1, pp. 759-772, 2024.

[3] Rieger A, Roth T, Sedlmeir J, Fridgen G, Young A, "Organizational identity management policies," Journal of the Association for Information Systems, vol. 25, no. 3, pp. 522-527, 2024.

[4] Oduri S, "Continuous authentication and behavioral biometrics: Enhancing cybersecurity in the digital era," International Journal of Innovative Research in Science Engineering and Technology, vol. 13, no. 7, pp. 13632-13640, 2024.

[5] Fan Y, Li J, "Cross-chain transaction of notaries based on two-stage improved PageRank algorithm," International Journal of Network Security, vol. 27, no. 1, pp. 95-103, 2025.

[6] Aguilera R V, Ruiz Castillo M, "Toward an updated corporate governance framework: Fundamentals, disruptions, and future research," BRQ Business Research Quarterly, vol. 28, no. 2, pp. 336-348, 2025.

[7] Rijanto A, "Blockchain technology roles to overcome accounting, accountability and assurance barriers in supply chain finance," Asian Review of Accounting, vol. 32, no. 5, pp. 728-758, 2024.

[8] Attaran M, "Blockchain technology in healthcare: Challenges and opportunities," International Journal of Healthcare Management, vol. 15, no. 1, pp. 70- 83, 2022.

[9] Li D X, Lu Y, Li L, "Embedding blockchain technology into IoT for security: A survey," IEEE Internet of Things Journal, vol. 8, no. 13, pp. 10452-10473, 2021.

[10] Panda S S, Jena D, Mohanta B K, Ramasubbareddy S, Daneshmand M, Gandomi A. H, "Authentication and key management in distributed IoT using blockchain technology," IEEE Internet of Things Journal, vol. 8, no. 16, pp. 12947-12954, 2021.

[11] Saputra I, Arkeman Y, Jaya I, Hermadi I, Akbar N A, Sutedja I, "AniraBlock: A leap towards dynamic smart contracts in agriculture using blockchain based key-value format framework," Communications in Science and Technology, vol. 8, no. 2, pp. 154-163, 2023.

[12] Saputra I, Arkeman Y, Jaya I, Hermadi I, Sutedja I, "Blockchain-based key-value store to support dynamic smart contract interaction in the agricultural sector," Indonesian Journal of Electrical Engineering and Computer Science, vol. 33, no. 1, pp. 622-633, 2024.

[13] Yang Y, Qiu Y, Cao J, Xiao S, "Identity-Based Blind Signature Scheme with Message Recovery over SM9," Chinese Journal of Electronics, vol. 34, no. 2, pp. 510-519, 2024.

[14] Lai J, Huang X, He D, Wu W, "Provably secure online/offline identity-based signature scheme based on SM9," The Computer Journal, vol. 65, no. 7, pp. 1692-1701, 2022.

[15] Wu Z J, Zhang Y, Yang Y M, Wang P, Meng Y, "BDSec: Security authentication protocol for BeiDou-II civil navigation message," China Communications, vol. 21, no. 6, pp. 206-218, 2024.

[16] Zhang X, Ye C, "A novel privacy protection of permissioned blockchains with conditionally anonymous ring signature," Cluster Computing, vol. 25, no. 2, pp. 1221-1235, 2022.

[17] Prabakaran D, Ramachandran S, "Multi-factor authentication for secured financial transactions in cloud environment," Computers, Materials & Continua, vol. 70, no. 1, pp. 1781-1798, 2022.

[18] Mokayed H, Quan T Z, Alkhaled L, Sivakumar V, "Real-time human detection and counting system using deep learning computer vision techniques," artificial Intelligence and Applications, vol. 1, no. 4, pp. 221-229, 2023.

[19] Gad A G, Mosa D T, Abualigah L, Abohany A A, "Emerging trends in blockchain technology and applications: a review and outlook," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 9, pp. 6719-6742, 2022.

[20] Mukherjee A, Balachandra M, Pujari C, Tiwari S, Nayar A, Payyavula S R, "Unified smart home resource access along with authentication using Blockchain technology," Global Transitions Proceedings, vol. 2, no. 1, pp. 29-34, 2021.

[21] Garba A, Chen Z, Guan Z, Srivastava G. "LightLedger: A novel blockchain-based domain certificate authentication and validation scheme," IEEE Transactions on Network Science and Engineering, vol. 8, no. 2, pp. 1698-1710, 2021.

[22] Anisetti M, Ardagna C A, Bena N, Damiani E. "Rethinking certification for trustworthy machine-learning-based applications," IEEE Internet Computing, vol. 27, no. 6, pp. 22-28, 2023.

[23] Miao J, Wang Z, Wu Z, Ning X, Tiwari P. A blockchain-enabled privacy-preserving authentication management protocol for Internet of Medical Things. Expert Systems with Applications, vol. 237, pp. 121329, 2024.

[24] Kumar R, Javeed D, Aljuhani A, Jolfaei A, Kumar P, Islam A N. Blockchain-based authentication and explainable AI for securing consumer IoT applications. IEEE Transactions on Consumer Electronics, vol. 70, no. 1, pp. 1145-1154, 2024.

[25] Kang T, Woo N, Ryu J. Enhanced lightweight medical sensor networks authentication scheme based on blockchain. IEEE Access, vol. 12, pp. 35612-35629, 2024.

Coordination, Communication and Robustness in Multi-Agents: An Industrial Network Scenario Using Trust Region Policy Optimization

Munam Ali Shah
Department of Computer Networks and Communication-College of Computer Science and Information Technology,
King Faisal University, Al-Ahsa, Kingdom of Saudi Arabia

***Abstract*—Numerous practical uses necessitate multi-agent systems, including managing traffic, assigning tasks, regulating ant colonies, and operating self-driving cars, and drones. These systems involve multiple agents working together, communicating and engaging with their surroundings to achieve the highest possible total numerical reward. Deep Reinforcement Learning (DRL) approaches are used to address these multi-agent applications. In many circumstances, the use of agents raise challenges to safety and robustness. To address these issues, we develop a DRL based system in which multiple agents in an industrial network scenario interact with the real-world environment and act collaboratively and cooperatively. In proposed model, several agents collaborate with one another to complete tasks and maintain a safe state. To take actions cooperatively and collaboratively of agents in accordance with the safety robustness of policies, we apply DRL algorithms such as proximal policy optimization (PPO) and Trust Region Policy Optimization (TRPO) algorithms and DRL approaches. We apply Curriculum Learning (CL) for their better performance and training. In this study, a reward structure is also proposed which help agents to maintain their safe state. Mean reward, policy loss, value loss, value estimate and safety robustness are analyzed as performance matrix in this study. The results shows that the policy adopted in the proposed model perform comparably better than the other policies.**

***Keywords*—Safety robustness; reinforcement learning; multi-agents; safe state; collaboration**

I. Introduction

In an experiment on a cat's behaviour done by Thorndike [1], reinforcement learning (RL) begins with the trial and error method. In machine learning, RL is a significant approach. It's a method of determining which action offers the most reward [2]. Real-time learning involves an agent that interacts with the learning environment, takes actions on the current state, and proceeds to the next state while striving to maximize the numerical reward. DRL, which merges Deep Learning and Reinforcement Learning [3] [4], is utilized to handle high-dimensional environments. To surpass the limitations of single-agent systems, multi-agent systems are necessary. The agents in multi-agent systems communicate with one another, interact with the environment, and aim to maximize the overall numerical reward [5]. Many real-world applications, such as traffic control, task distribution, ant colonies, autonomous cars, and drones, necessitate multi-agent systems. Due to the complexity of a high-dimensional environment, an agent that is hard-coded cannot work well in it.

To ensure the agents can generate optimal learning solutions, Multi-Agent Deep Reinforcement Learning (MADRL) is a crucial technique in the previously mentioned applications. In these applications, MADRL agents can discover the most effective solutions and policies. MADRL is a learning method where multiple agents interact with the learning environment to maximize the total reward. Using neural networks with numerous layers and approximations, agents strive to estimate the value of actions. Their policies rely on their own action values and those of other agents [5].

Therefore, researchers are now focusing not only on reward maximization but also on safety. Deep integration of machine learning and control technologies is required to tackle safety challenges in critical applications such as self-driving cars and robot manipulation [4]. In the control community, robust and adaptive control techniques have been developed to ensure the safety and stability of systems. To enhance the safety of learned policies, appropriate control approaches from the control community can be applied [4] [6]. The performance of learned policies is typically measured through success rates [7] [8] or obtained rewards [9] [10] to determine their robustness. However, when using RL to address problems in safety-critical applications, policy robustness must also be evaluated from a safety perspective. However, there is presently no way of evaluating policy robustness from the standpoint of safety.

This work proposes a framework for the safety robustness of agents using deep reinforcement learning techniques. In the proposed work, a system is developed in which multiple agents interact with the real-world environment and act collaboratively and cooperatively. This work is an extension to our previously published work [33]. Here, we have added policy loss, safety robustness, value loss and value estimate. We have evaluated these parameters on multiple policies, i.e., Policy 1, Policy 2 and Policy 3. We have also proposed a novel algorithm for the training process. In proposed model, several agents collaborate with one another to complete tasks and maintain a safe state. To take actions cooperatively and collaboratively of agents in accordance with the safety robustness of policies. We have applied DRL algorithms such as proximal policy optimization (PPO) and Trust Region Policy Optimization (TRPO) algorithms and DRL approaches.

The rest of the paper is organized as follows: Section II provides the literature review. Section III presents the proposed methodology. In Section IV, the details of the proposed model are presented. Section V explains the implementation setup. In Section VI, experiments and results are presented and the paper is concluded in Section VII.

II. Literature Review

A. Single Agent Scenarios

Alternative reward functions may be presented to help an agent learn better if a reward function for an RL problem is not provided. With multiple reward functions, the agent may learn a variety of policies with varying levels of robustness. Evaluating the achieved reward and success rate has been a common method for assessing the robustness of a policy. When utilizing RL to address issues in many safety-critical situations, safety is a concern. Although reward and success rate have been commonly used to evaluate policy robustness, there has been limited discussion on assessing policy robustness from a safety perspective. This study [11] contributes to the field by introducing a novel concept of safety robustness and an algorithm to estimate it. The proposed method is applied to evaluate the safety robustness of three policies for controlling the manipulation of a cable-driven parallel robot. The results demonstrate that the algorithm is effective in estimating safety robustness by comparing the number of safe episodes to the total number of episodes. The study also includes illustrative experiments to demonstrate the implementation of the proposed algorithm. It can also choose the safest policy from a set of many policies.

As a recent and popular research area, RL faces a number of issues. Agent navigation and obstacle avoidance are the most appealing of these RL challenges. It is well understood the value of precise and multidimensional monitoring in detecting problems early and preventing them from escalating. Previously, AI was employed for drone navigation, with images from on-board cameras being processed for wayfinding and collision avoidance. The authors of this paper presented a novel recommender system for drone navigation, which utilizes artificial intelligence and sensor data while requiring only basic information [12]. Their approach involves combining two techniques: Proximal Policy Optimization (PPO) for navigation with minimal information and Long-Short Term Memory (LSTM) networks for overcoming obstacles using navigation memory. The safety requirements of the system are determined using a systematic functional failure analysis (FFA).

RL methods have been used to handle a variety of sequential decision-making issues for a long time. In complex environments with high dimensions, algorithms face significant challenges when attempting to optimize policies. Recent advances in deep learning have enabled RL approaches to drive optimal policies for complex agents capable of performing well in these difficult contexts [13] [14]. In their paper, the authors conducted a survey on various approaches for multi-agent deep RL (MADRL). The survey covers non-stationary, partial observability, continuous state and action spaces, multi-agent training methods, and multi-agent transfer learning [15]. The survey also discusses the advantages, disadvantages, and applications of the evaluated approaches, leading to the creation of more robust and practical multi-agent learning methods for tackling real-world issues in the future.

To overcome challenges in dealing with local-stable-point scenarios in complex situations, the artificial potential field method needs to be revised, which may increase the algorithm's complexity. In this study, the authors proposed an improved black-hole potential field combined with RL [16] to address this issue. The black hole potential field is used as the environment in an RL system, where agents automatically adapt to their surroundings and learn how to discover targets using basic environmental data. Furthermore, with the curriculum learning technique, taught agents adapt to a variety of contexts. Meanwhile, the avoidance process is visualized to show how agents avoid obstacles and achieve their destination. Static and dynamic trials are used to assess our approach. The findings indicate that agents can autonomously learn how to escape local stability points even without prior knowledge.

The authors presented a collision avoidance system in their research paper, which aims to provide reliable autonomous navigation for self-driving ships. The system uses a collision risk assessment [17] technique based on the ship's domain and the closest point of approach (CPA) to estimate the collision risk, and generates an avoidance path only when necessary. The ship domain is designed with an asymmetric shape, taking into account the ship's maneuverability and the international regulations for preventing collisions at sea (COLREGs). The CPA is used to determine the value of a quantitative collision risk. The authors of [18] investigate the resilience of RL with adversarial altered state observations, which is relevant for deploying real-world RL agents in the face of unexpected sensing noise. They show that with a given agent strategy, an optimum adversary to tampering state observations can be discovered, ensuring that the worst-case agent reward is obtained. This results in a unique empirical adversarial assault on RL agents in DRL environments, via a learnt adversary that is significantly stronger than earlier ones. Authors offer an alternate training with learned adversaries (ATLA) approach to improve an agent's resilience by training an opponent online alongside the agent using policy gradients and the optimum adversarial attack framework.

The automobile industry is being encouraged toward more flexibility and stability in manufacturing by EU laws on CO2 limitations and the trend of individualization. Modular manufacturing, in which workstations are detached by autonomous guided vehicles and new control ideas are required, is one way to handle these issues. Throughput-optimal coordination of goods, workstations, and vehicles are the goals of modular production control. Conventional control techniques for this NP-hard issue are inefficient in computation, do not discover optimal solutions or are not generalizable. On the other hand, DRL provides strong and generalizable algorithms that can deal with a wide range of settings and high levels of complexity. The PPO method is utilized in solving modular production control, as discussed in the paper [19]. The experimental results showed that the learned behavior of the agent was optimal, reliable, consistent, and generalizable across various modular production control setups. The agent was successful in adapting its strategies based on the given problem configuration. The authors further discussed how this learning behavior was achieved, with a focus on the agent's state, activity, and incentive design.

B. Multi-Agents

The robots that work in the marine environment are Unmanned surface vehicles (USV), which is a surface ship that can detect the target, and do the perception of the environment.

TABLE I. IN THE LITERATURE REVIEW TABLE, THE FEATURE THAT HAS NOT BEEN ADDRESSED IN PREVIOUS WORK IS DENOTED BY ×, WHILE √ INDICATES THAT THE FEATURE HAS BEEN ADDRESSED IN PRIOR RESEARCH

Ref.	Multi Agents	Agents make joint actions	Single Agent	Partially observable environment	Real world complex environment
[5]	√	√	√	×	×
[11]	×	×	√	×	×
[20]	×	×	√	×	√
[12]	×	×	√	×	×
[29]	×	×	√	×	×
[30]	√	×	×	×	×
[34]	×	×	√	×	×
[35]	√	×	×	×	×
Proposed Model	√	√	√	√	√

Autonomous navigation and obstacle avoidance are very necessary for USV and have been frequently employed to carry out operations in a complex sea environment or in risky marine areas for ships with sailors, considerably increasing their defensive capability and detection range [26] [27]. Traditional navigation approaches increase the risk of falling into traps while travelling in a complex, dynamic environment, and the likelihood of reaching the target decreases. These heuristic approaches are slow, and when used in real-world sophisticated dynamic situations, they have issues such as slow speed and inability to identify and avoid obstacles [24] [25]. Authors proposed a DRL approach called ANOA (Autonomous Navigation and Obstacle Avoidance), which uses a dueling deep Q-network and a customized architecture of state and action spaces [20]. They implemented their proposed approach in a static and dynamic environment using Unity 3D and the ML-Agents Toolkit. Their experimental results show that once they are trained, USVs can find a path, reach their destination, and avoid obstacles in any complex dynamic environment. In [21], authors proposed a DRL based framework for the collaboration and cooperation of multi agents in real-world environments. A partitioned base technique is presented in this work. For effective coordination and continuous operation of a multi-agent system, it is necessary for agents to have knowledge of the environment's boundary and the location of the currently active neighboring agent. This allows for seamless transition to the next neighboring agent in case of any unexpected issue or failure.

Authors explore the feasibility of utilizing the PPO, a cutting-edge DRL algorithm for continuous control tasks, to address the challenge of effective path planning that caters to the two primary objectives of path following and collision avoidance (COLAV), essential for ensuring the safety and dependability of autonomous surface vehicles (ASV) [29]. The AI agent is trained and evaluated in a complex, stochastically generated simulation environment employing the OpenAI gym Python toolbox equipped with several rangefinder sensors for obstacle detection. The agent has real-time visibility into its own reward function, which enables it to adapt its guidance strategy dynamically. The trained agent achieves an episodic success rate of almost 100 percent, varying from path-adherence to obstacle avoidance.

The application of deep neural network-based systems has become a state-of-the-art approach in various robotics tasks. However, utilizing these systems in safety-critical areas can be precarious in the absence of explicit network robustness guarantees [22] [23]. Small changes in sensor inputs (due to noise or adversarial situations) can cause network-based decisions to change, as demonstrated by an autonomous car swerving into another lane. To mitigate the risks posed by adversarial attacks, various defensive algorithms have been developed, including those that provide formal robustness guarantees. In the study [30], certified adversarial robustness research is utilized to develop an online DRL method that is certifiably robust. The proposed defense mechanism calculates lower bounds on state-action values to detect and select a resilient action in case of worst-case input divergence caused by possible adversaries or noise. Additionally, the resulting policy is accompanied by a certificate of solution quality, despite the certifier's lack of knowledge about the real state and optimal action due to disturbances. The study provides new performance guarantees, extensions to other RL algorithms, aggregated results across multiple scenarios, an extension into scenarios with adversarial behavior, comparisons with a more computationally expensive method, and visualizations that provide intuition about the robustness algorithm [31] [32].

In Table I, the existing literature is comprehensively summarized. They must guarantee that DRL-trained robots are robust in order for them to function alongside humans in actual circumstances. Because the current world is so diverse, and human behaviour frequently changes in reaction to agent deployment, the agent will almost certainly encounter scenarios that he or she has never encountered before. This creates an assessment challenge: how can authors successfully measure robustness if they can't use the average training or validation reward as a metric? Unit testing in software engineering is a source of inspiration for them. In their paper [34], the authors propose that when creating AI agents that interact with people, designers should consider potential edge cases in partner behavior, conceivable states encountered, and construct tests to guarantee that the agent's behaviour is appropriate in these instances. This technique is used to create a suite of unit tests for the Overcooked-AI environment, which is then used to assess three ideas for enhancing robustness. Authors discover that using the test suite gives them a lot of information about the consequences of these proposals that they couldn't get by just looking at the average validation reward.

III. JUSTIFICATION OF THE PROPOSED MODEL

It would be naive to assert that MAS should be employed in all complex system designs. There are certain cases when it is especially suited, as with any beneficial strategy, and others where it is not.Multi-Agent Systems (MAS) have numerous real-world applications, including but not limited to traffic control, task distribution, ant colonies, autonomous cars, and drones. A hard-coded agent cannot operate well in a high-dimensional environment due to the complexity of the environment. MADRL agents can find the best solutions and policies. The policies of other agents are impacted by changes in the environment when agents collaborate and coordinate their actions. Failure of a learnt policy in safety-critical applications might result in dangerous scenarios. The purpose of this section is to emphasise the importance and utility of MADRL in accordance with safety. In safety-critical

applications, ensuring the robustness of policies is essential when using RL to solve problems. To enhance the safety of learned policies, control approaches from the control community can be utilized. Safety-critical scenarios are not considered in multi-agent settings. So, there is a need to develop a DRL-based system that tackles the problem of safety in a multi-agent environment.

IV. Proposed Model

A technique for approximating the safety robustness of policies is proposed, which is based on the system architecture of robust control, safe states (i.e., non-error states [35]), and statistical analysis. The safety robustness is approximated by calculating the ratio of the number of safe episodes to the total number of episodes, which can be used as an indicator to evaluate policy robustness from a safety perspective. This model uses curriculum learning for the randomization of the environment and the better performance of agents. Agents start their learning with easy tasks and then move to the difficult ones. We use PPO to analyze the visual observations as well as vector observations. By using visual and vector observations, agents communicate with each other. In order to train our model, we employ a 4-dimensional action space in the environment. CNN is the neural network used by PPO (a fully linked neural network).

The environment is sensed by the agent's sensors, which are subsequently passed on to the first layer of the neural network. The activation function is used here before moving on to the next layer. Similarly, the agent receives the output after using the activation function. The agent receives this output and estimates his action based on that output. The agent receives a positive reward if it takes the correct action, while a negative reward is given if it takes the wrong action.

We also use CL (a method of training a machine learning model that gradually introduces more challenging components of a problem such that the model is constantly optimally challenged) for training purposes also.

A. Proposed Method of DRL

When agents interact with the environment, there are a number of observations made by them. According to observations, an agent must follow some policy and, as a result, must take action [28]. By taking that action, the agent makes different changes in the environment. Changes may be positive (perform the best action to tackle the task) or may be negative. (perform bad actions, exploit the task or environment). After performing an action, the agent moves to another state. The agent again observes the environment and follows another policy to gain the optimal policy for that task.

We employ the PPO algorithm for training in our multi-agent settings, and by combining it with CL, we can expand the training process to a large number of agents. The learning environment is entirely cooperative, and each agent possesses only basic knowledge of the surroundings. The goal of all agents is to maximize the total discounted reward by working together. The following is a description of a tuple:

$$(N, S, A_i, Z_i, C, R, O)$$

N denotes the total number of agents in the system, S represents the number of possible states that the system can be in, and A_i denotes the number of possible actions that agent i can take. C and R represent the joint transition and reward functions, respectively. O represents the number of possible observations that can be made, and Z_i represents the collection of observations that agent i has access to.

To make the training more stable, each agent only knows limited knowledge about the environment in a single state S_c. As a result, agents must recall some information from prior observations.

The utilization of the PPO algorithm was employed to improve training with limited rewards. This method involves the use of a conditional statement to differentiate between actions and observations derived from the agent's demonstration. The condition assesses whether the new actions and observations are similar to those in the demonstration. The agents are rewarded for their efforts. We may also take advantage of local incentives using PPO. This investigation employs an extrinsic reward signal as a local reward signal, and utilizes a combination of CL with RL and PPO to expedite policy convergence. Fig. 1 is the demonstration of the proposed system model.

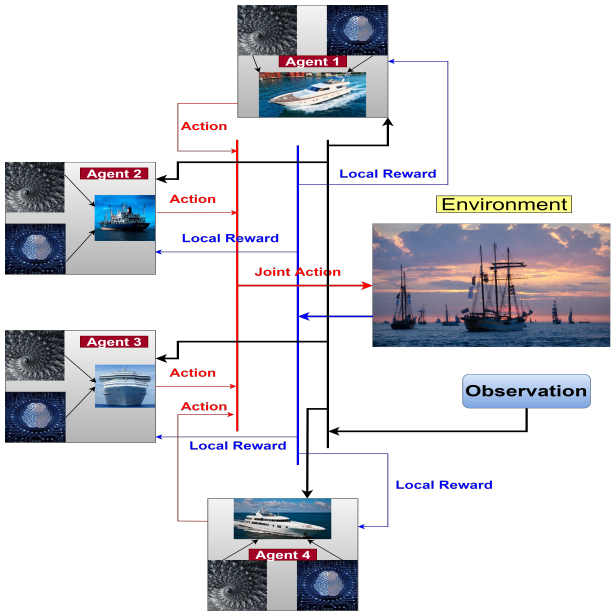


Fig. 1. Block diagram of the proposed system model [33].

In Fig. 2, obtained from [33], you can see that agents observe visual and vector observations from the environment and move these inputs (observations) to the convolutional layer. The agent receives the outputs of the results. According to these outputs, agents take action in the environment.

The mentioned techniques are implemented for each agent, which operate by collaborating and cooperating with other agents. The agents interact with the learning environment through cooperative actions, and as a result, they receive a local cumulative reward.

This allows agents to quickly imitate demonstrated behaviour, which speeds up the training process. The CL tech-

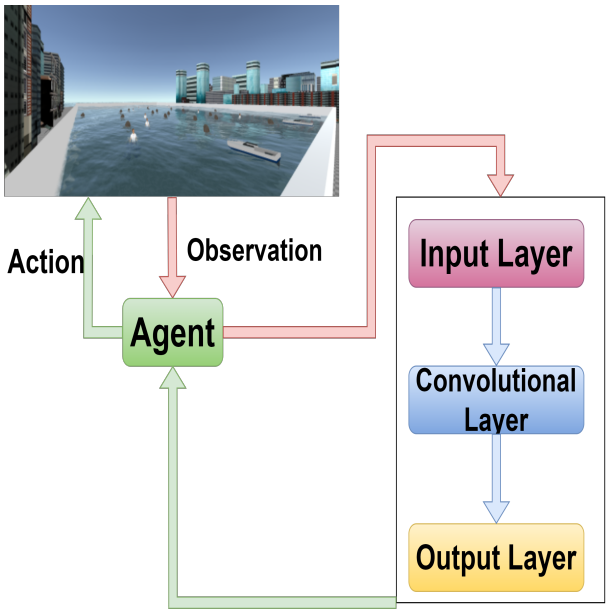


Fig. 2. Layered approach in the proposed system model.

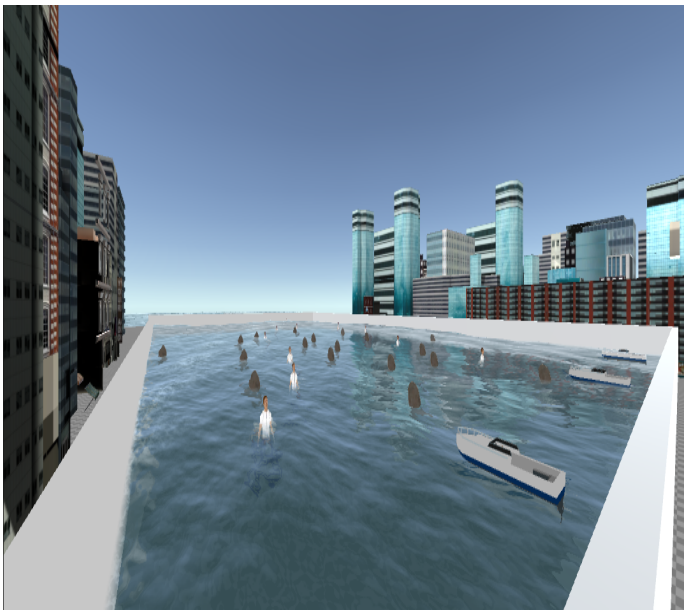


Fig. 3. A scene of the learning environment for agents.

nique guarantees that the policies are converged quickly. The algorithm provides the pseudo-code for the suggested model.

B. Components of the Proposed Model

Different components of proposed model are as under:

1) Learning environment: In many aspects of life, robots play an important role. Recently, robots have been used in digital factories [36], as well as in hazardous locations such as fires, storms, and landslides. Robots can carry, move, and lift large things from one location to another, making it easier to provide goods to people who are in risky or challenging situations. Robots operate more effectively and quickly than people in difficult situations, reducing human risk. These well-trained agents have also been utilised in the marine environment to rescue individuals and apprehend smugglers. As a result, we decided to learn in a maritime setting. This paper looks at a real-world situation in which agents try to save people by getting around problems and working together.

The learning environment employed in this study is depicted in Fig. 3. Agents in a marine setting navigate the goal (people in need of rescue) and navigate to them while avoiding obstacles. For example, the environment contains obstacles in the form of heavy rocks, and the agents have the ability to move in four directions: left, right, forward, and backward. Agents go towards their target by doing these activities and avoiding obstacles through navigation and coordination with other agents.

2) Agents: In Unity 3D games, objects are used as agents in the learning environment. The agent is a boat that is used to move in a marine environment and to rescue individuals who are drowning. In Fig. 4 agent is demonstrated as "ship".

In the learning environment, agents generate states and execute actions from a set of predefined actions such as left, right, forward, and backward, thereby transitioning from one state, denoted by S_t to the next state, denoted by S_{t+1}. Each agent is associated with a single brain that guides the agent on its next course of action.

Fig. 4. A scene of an agent navigating in the given environment.

C. Training Process Algorithm

In Algorithm 1, it is observed that the process starts from episodes. Here we apply a "For-loop" that initiates the training process from the first episode.

Algorithm 1 Training Process

For episodes = 1 to n
{
Initialize the Environment
Randomize the Environment
For Time-steps = 1 to 5000
{
For each Agent $a_i < 4$
Agent a_i observe Environment
Calculate distance of every agent
{
if (Distance <= 50 meters of Agent a_i AND Agent a_j)
{
Broadcast message to Agent a_j
}
if (Goal of Agent a_i AND Agent a_j == same)
{
Agent with minimum distance from goal will take action.
}
}
Calculate Reward.
}
Mean Reward = *Actions with positive reward / Total Reward*
if (Mean Reward > 0)
{
Safe Episode
}
}
Safety Robustness of Current Policy = *total Safe Episodes / total Episodes*

Episode and then moves on to the next to meet the desired number of episodes. The environment is initialised in each episode. After initialization, the environment is randomized. Because the examples of dynamic environments are considered so, there is a need to observe all the environments. After randomizing the environment, another "For-loop" is applied inside the first for-loop. This for-loop starts from the first time-step to the 5000 time-steps of each episode. Within this loop, there is an additional nested "for-loop" that iterates through each agent in the environment.

In each time-step, every agent observes the environment and calculates their distance from other agents. Here is the if statement. If the goals of both agents are common, then the agent having the minimum distance from the goal will take action and collect a reward while the other agent will select the other target/goal. If both agents have different goals, both agents will perform or take their actions according to their targets. When 1 episode is completed, the mean reward will be calculated by dividing the total positive reward by the total reward. The following "if condition" checks if the mean reward is above zero. If this condition is satisfied, the episode is considered safe. After all episodes are completed, the policy's safety and robustness are determined by calculating the ratio of safe episodes to total episodes.

D. Flow Chart

The flow chart of our proposed model is shown in Fig. 5. It is observed that in each episode, the environment is initialized. Because after initialization, the environment is randomized. because the examples of dynamic environments are considered. Here is a loop. If time steps are completed, the process will start from another episode. If the time-steps are pending, then in each time-step, every agent observes the environment. According to that observation, agents calculate their distance from other agents. After calculating the distance, agents having a minimum distance of 50 metres broadcast their goals. If the goals of both agents are common then the agent having minimum distance from the goal will take action and collect reward and the other agent will select the other target/goal. If both agents have different goals, both agents will perform their actions to meet their target. After completing one episode, the mean reward can be calculated by dividing the total positive rewards obtained during the episode by the total number of rewards received. Completion of all episodes results in the safety and robustness of the current policy.

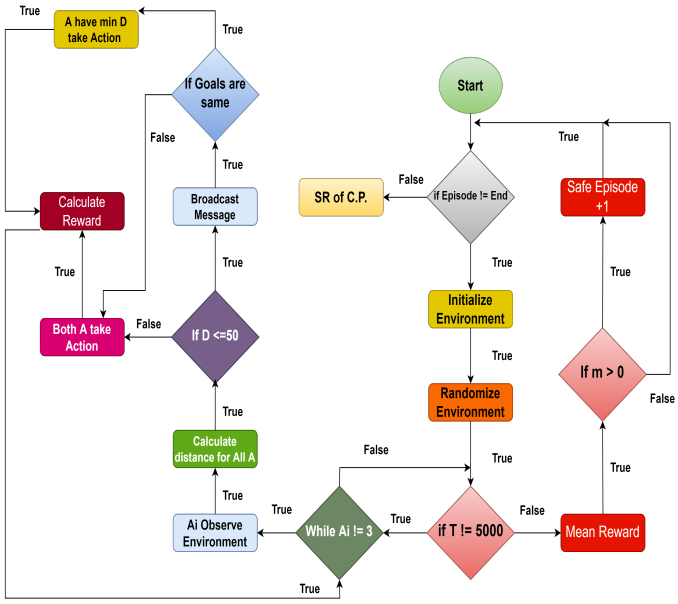


Fig. 5. The flow chart of the proposed system.

E. Reward Function

To promote desirable behavior, a local incentive signal is utilized to encourage consistency. An effective incentive system ensures faster convergence of the learning process in single-agent contexts than ever before [37]. The entire reward for a single agent may be calculated using the equation below:

$$r = (r_{hp} + r_{hc}) - (r_{cla} + r_{clo} + r_{cldo})$$

where, r is the sum of rewards given to each agent, where r_{hp} is the reward for rescuing drowning humans, r_{hc} is the reward for rescuing all humans in the environment, r_{clo} is the negative reward for colliding with an obstacle, r_{cla} is the negative reward for colliding with other agents and (r cldo) is the negative reward for colliding with some dynamic or sudden obstacle. r_{hp} can be calculated using the following equation:

$$r_{hp} = \begin{cases} r_p & \text{if} \quad d = 0 \\ 0 & \text{Otherwise} \end{cases}$$

where, r_p is the reward for rescuing people and 'd' represents the distance between the agent and the drowning person.

Moreover, we can calculate reward r_c using the following equation:

$$r_{hc} = \begin{cases} r_c & \text{if} \quad h = 0 \\ 0 & \text{Otherwise} \end{cases}$$

where, h_c is the reward of rescuing all drowning people, and h is the number of drowned people remaining in the sea.

Furthermore, to calculate the r_{cla} reward, we can use the following equation:

$$r_{cla} = \begin{cases} r_a & \text{if} \quad d = 0 \\ 0 & \text{Otherwise} \end{cases}$$

the reward when an agent collides with another agent is denoted by r_a and d is the distance between the two agents.

Moreover, to calculate the r_{clo} reward, we present the following equation:

$$r_{clo} = \begin{cases} r_o & \text{if} \quad d = 0 \\ 0 & \text{Otherwise} \end{cases}$$

where, r_o is the reward given to the agent on collision with static obstacles, and d is the distance between the agent and the static obstacle.

Lastly, we present the equation to calculate the r_{cldo} reward.

$$r_{cldo} = \begin{cases} r_{do} & \text{if} \quad d = 0 \\ 0 & \text{Otherwise} \end{cases}$$

where r_{do} is the reward of an agent's collision with a moving drowning human or any dynamic obstacle, and d is the distance between the agent and the drowning human or dynamic obstacle.

The table below illustrates the awards each agent received for various occurrences. We implemented a penalty of -0.1 for each agent when they collided with another agent in the environment. This approach addresses the issue of inter-agent collision. We also assign a value of -0.1 when the agent collides with the seashore or any other obstruction. The agent earns-0.1 reward if it collides with any drowning human in the surroundings. An agent who successfully rescues a drowning individual receives a +0.3 bonus. Furthermore, the agents will gain a +0.5 bonus if they are successful in rescuing all drowning people in the surroundings.

V. Implementation

This section describes how we will put the suggested approach into practice. We also go through the tools that were utilised to carry out the implementation. This paper also investigates various behavioral characteristics and training parameters.

A. Unity 3D

Unity Technologies is the developer of the cross-platform game engine called Unity, which was first announced and released as a game engine for Mac OS X in June 2005 at Apple Inc.'s Worldwide Developers Conference. Over time, the engine has been enhanced to support various platforms such as PC, mobile, console, and virtual reality. It allows developers to create 3D and 2D games, interactive simulations [38], and other experiences.

Gazebo is also a three-dimensional dynamic robot environment simulator that can be used to test algorithms and robot designs. It runs regression tests and uses realistic scenarios to teach AI systems. The speed of the Gazebo Simulator has never been a strong point. Simulating a 30-minute event takes 6 hours. Multiple instances operating at the same time is tough. We can make it happen, but there are still difficulties like non-accuracy and an unknown amount of effort required to coordinate among these various instances.

Unity [39], on the other hand, is far more efficient at running several simulations on a single system and coordinating learning across all instances. Unity simulations may be cranked up to a hundred times faster than real time.

B. ML-Agents Toolkit

The toolbox includes numerous components for implementing RL, such as Brain, Academy, and Agent. Each agent is connected to and controlled by a brain component. Different environmental factors are learned using the Academy component. Additionally, it provides a link between the brain and Python Tensorflow. Fig. 6 shows the basic structure and overview of the ML-Agents toolkit [40].

C. Training Parameters

The learning environment's working area is 200×200 square feet. The functional area of the learning environment contains a variety of objects, including rocks, boats, humans, and agents. The combination of moving individuals and robots was utilized to enhance the realism and dynamism of the learning environment, enabling the agents to be trained in such a setting.

TABLE II. Training Parameters Used in the Proposed Model

Parameters	Values
Batch size	1024
Buffer size	16384
Learning rate	0.0003
Max steps for each episode	5000
Number of epoch	3
Number 0f layers	2
Gamma in extrinsic Reward signal	0.99
Behavior cloning strength	0.5

Fig. 6. ML-Agents toolkit used in the proposed model.

Table II displays some important training parameters employed to train the agents utilizing the proposed model. Based on empirical testing, a batch size of 1024 is deemed appropriate. Additionally, the buffer size parameter must be larger than the batch size. The buffer size was set to 16384 by default. Local RL reward signals are used in the proposed model.

TABLE III. TRAINING PARAMETERS USED FOR POLICY 1

Parameters	Values
Batch size	1024
Buffer size	16384
Learning rate	0.0003
Max steps for each episode	5000
Number of epoch	3
Number 0f layers	2
Gamma in extrinsic Reward signal	0.99
Time horizon	128
Hidden Units	512

The Table III provides some key training parameters for policy 1, which were utilized to train the agents. The batch size for the PPO algorithm has been adjusted to 1024. By default, we set the buffer size parameter to 16384, which is greater than the batch size of 1024. For policy 1, we set the temporal horizon value to 128 and the hidden units value to 512.

The Table IV provides some key training parameters for policy 2, which were utilised to train the agents. The batch size for the SAC algorithm has been adjusted to 1024. By default, we set the buffer size parameter to 16384, which is greater than the batch size of 1024. For the SAC algorithm, we set the temporal horizon to 128 and the hidden unit value to 512. Initial buffer steps of 6000 are used. Tua is used at a value of 0.01. For policy 2, we employ 10.0 steps for each update.

TABLE IV. TRAINING PARAMETERS USED FOR POLICY 2

Parameters	Values
Batch size	1024
Buffer size	16384
Learning rate	0.0003
Max steps for each episode	5000
Number of epoch	3
Number 0f layers	2
Gamma in extrinsic Reward signal	0.99
Time horizon	128
Hidden Units	512
Tau	0.01
Buffer initial steps	6000
Steps per update	10.0

VI. EXPERIMENTS AND RESULTS

With the aforesaid configuration, Algorithm 1 is used to evaluate the safety robustness of the three approaches for managing USVs in terms of obstacle avoidance and navigation to their destination. Fig. 7 illustrates the orientation of the USVs controlled by the three policies (p1, p2, and p3) towards a goal orientation when the maximum uncertainty of action and state is 0.1. It can be observed that all three policies successfully maintain the USVs in safe states.

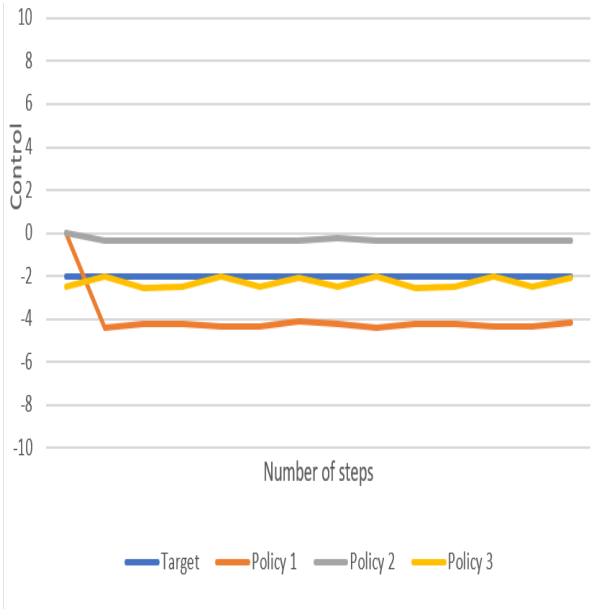


Fig. 7. Orientation of the USVs controlled by the three policies with action uncertainty value 0.1.

The y-axis depicts the control of USVs, while the x-axis represents the number of time-steps.

The target is shown by a blue line, the policy-1 is indicated by a red line, the policy-2 is indicated by a grey line, and the policy-3 is indicated by a yellow line.

The proposed model's policy is policy-3, and it can be observed that policy-3 is closer to the aim than the other policies. It can be seen from Fig. 7, the policy followed by our proposed model is more resilient in terms of USV's control than the other policies.

When the maximum magnitudes of the uncertainties of action and state are 0.2, Fig. 8 shows the results.

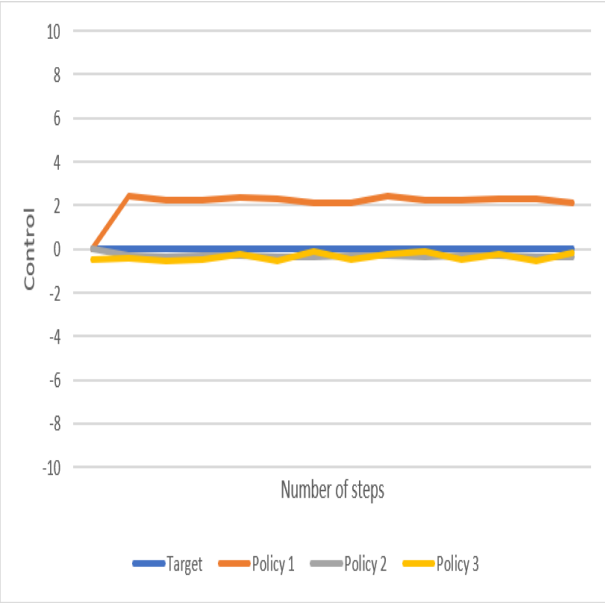


Fig. 8. USV control with action uncertainty of 0.2.

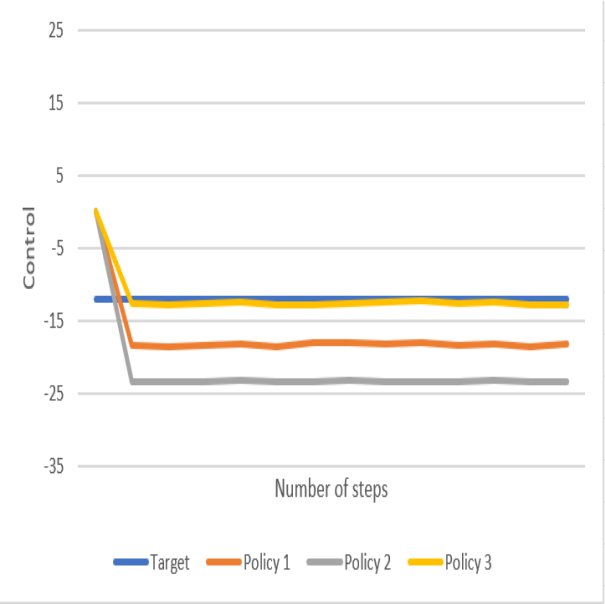


Fig. 9. USV control with action uncertainty of 0.5.

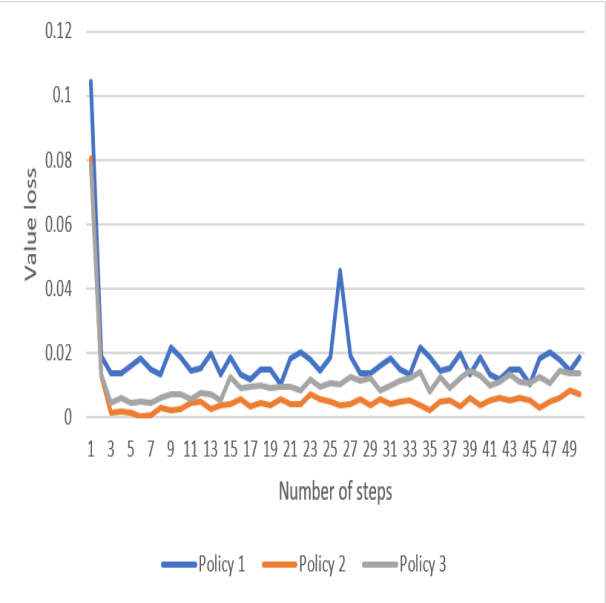


Fig. 10. Value loss against the number of steps in 3 different policies.

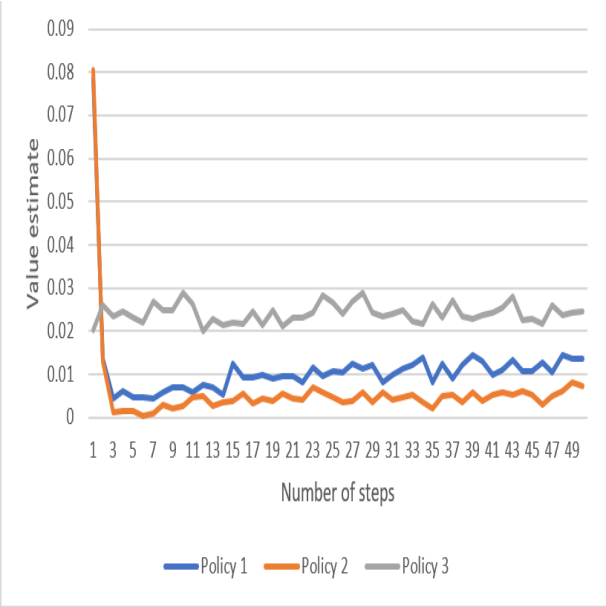


Fig. 11. Value estimate against the number of steps in 3 different policies.

The USV control is represented on the y-axis, while the number of time-steps is represented on the x-axis.

The plot displays the target as a blue line, while the policy-1, policy-2, and policy-3 are represented by a red, grey, and yellow line, respectively.

The proposed model's policy is policy-3, and it can be observed that policy-3 is closer to the aim than the other policies.

It can be seen from Fig. 8, the policy followed by our proposed model is more resilient in terms of USV's control than the other policies.

Fig. 9 illustrates the orientation of USVs controlled by the three policies (p1, p2, and p3) to a goal orientation, when the maximum magnitudes of the uncertainty of action and state are 0.5. All three policies were able to maintain the USVs in safe states.

The y-axis depicts the control of USVs, while the x-axis represents the number of time-steps.

The blue line represents the target, while the red, grey, and yellow lines represent policy-1, policy-2, and policy-3, respectively.

The episode is stopped at the 40th step when the USV is controlled by p1 since it is out of safe condition. The USV

can be kept safe by policies p2 and p3.

As a result, the three policies have varying degrees of ability to keep the USV in a safe state. Before the policy with the best safety robustness, it is crucial to evaluate the safety robustness of the three policies using Algorithm 1.

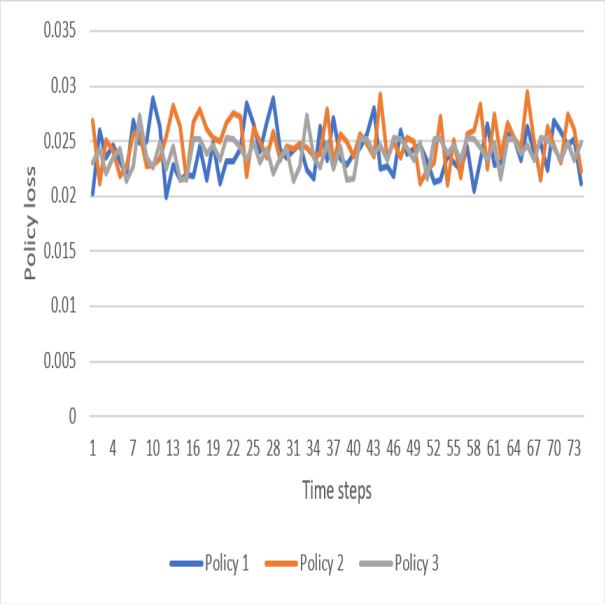


Fig. 12. Policy loss against the number of steps in 3 different policies.

In Fig. 10, the y-axis represents the value loss, while the x-axis indicates the number of time-steps. The blue line represents policy-1, the orange line represents policy-2, and the grey line represents policy-3. The graph indicates that policy-3 has an intermediate value loss between policy-1 and policy-2, suggesting that policy-3 has an average value loss.

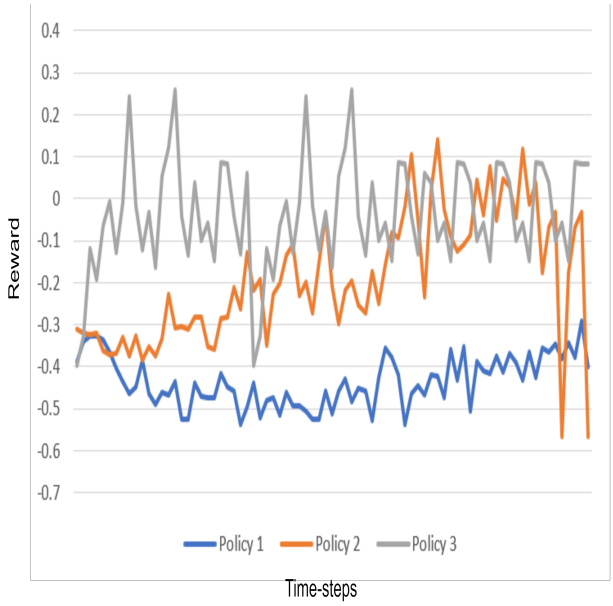


Fig. 13. Reward against the number of steps in 3 different policies.

Fig. 12 plots the value loss against the number of time-steps, where the y-axis represents the value loss, and the x-axis represents the time-steps. The performance of the three policies (i.e., policy-1, policy-2, and policy-3) is shown in blue, orange, and grey lines, respectively. The figure demonstrates that policy-3 has the lowest value loss among the three policies.

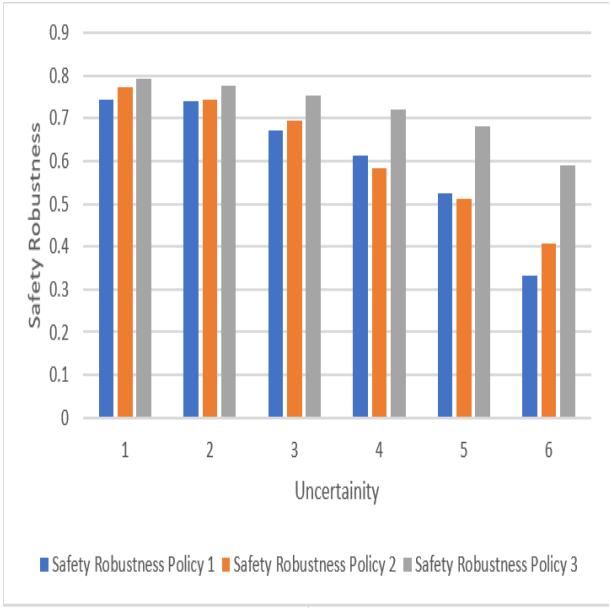


Fig. 14. Safety robustness vs. Uncertainty in 3 different policies.

The plot in Fig. 11 shows the value loss on the y-axis against the number of time-steps on the x-axis. The three policies, Policy 1, Policy 2, and Policy 3, are represented by blue, orange, and grey lines, respectively. It is evident from the plot that Policy 3 has the highest policy loss among the three.

The graph in Fig. 13 illustrates the rewards obtained by all policies, where Policy 1 is represented by a blue line, Policy 2 by an orange line, and Policy 3 by a grey line. The reward is a measure of policy efficiency, and the policy with the highest reward indicates better efficiency. From the graph, it is evident that Policy 2 gains more rewards compared to Policy 1, and Policy 3 gains more rewards than both Policy 1 and Policy 2. The proposed model adopts Policy 3, which is demonstrated to be the most efficient among the three policies.

In Fig. 14, the y-axis represents the reward gained by all policies, while the x-axis indicates the number of episodes. The blue line represents Policy-1, the orange line represents Policy-2, and the grey line represents Policy-3. The safety robustness of a policy indicates its ability to keep the system in a safe state.

Efficiency of a model is indicated by its safety robustness, and the policy that has maximum safety robustness is considered the most efficient. In Fig. 14, it can be observed that the proposed model follows policy 3, which has the highest safety robustness among all policies.

VII. Conclusions

In this work, we investigated a DRL based scenario which use multi-agents. These agents have been assigned some tasks

related to real world scenario. In our case, a submarine scenario was considered and the task of the agents is to safe the drowning person. We extended our previously published work by adding more experiments such as policy loss, reward function and the safety robustness. While the training of agents using deep reinforcement learning in real-world scenarios is very important, the safety of agents and other human beings is also very important. It is very important to learn about a policy that is robust in the context of safety. Several previous works in single agent settings focused on the safety of agents and other aspects of their training environment. But no work was done on the robust safety of agents in multi-agent settings. This work is focused on the safety and robustness of policies in multi-agent settings. A model is proposed that ensures the safety of agents and other valuable entities in the environment. The proposed model uses PPO algorithm to train the agents. Curriculum learning is used to teach the agent about safety and find a robust policy from easy to difficult level. A sea environment is considered in this work where some USVs are rescuing human beings that have been dropped into the sea. An algorithm is proposed in this research for the training of agents and finds a robust policy in the context of safety. To evaluate the safety robustness of policies, a few parameters are evaluated in this work. Safety robustness with respect to the number of safe states and control of USVs for the target with respect to the number of steps is evaluated, which shows the robustness of policy in safety. The policy followed by the proposed model is compared with the other two policies. Through experimental evaluation, it has been observed that the proposed model outperforms other policies. Furthermore, the performance of the proposed model has been verified by evaluating parameters such as mean reward, policy loss, policy estimate, and value loss. The experimental results have demonstrated that the proposed model exhibits better performance in multi-agent scenarios.

In the future, we aim to test our model with a higher number of agents in more complex scenarios. For example, it is vital to investigate how wind and its speed can effect the safety and robustness of the agents in the DRL scenario.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support from Department of Neteorks and Communications, College of Computer Science and Information Technology, and Deanship of Scientific Research (DSR), King Faisal University, Al-Ahsa, Kindom of Saudi Arabia.

REFERENCES

[1] E. L. Thorndike, Animal intelligence: An experimental study of the associate processes in animals., American Psychologist 53 (10) (1998) 1125.

[2] etelli, M. Pirotta, M. Restelli, Compatible reward inverse reinforcement learning, in: The Thirty-first Annual Conference on Neural Information Processing Systems-NIPS 2017, 2017.

[3] K. Arulkumaran, M. P. Deisenroth, M. Brundage, A. A. Bharath, Deep reinforcement learning: A brief survey, IEEE Signal Processing Magazine 570 34 (6) (2017) 26–38.(2020) 230–240.

[4] N. D. Nguyen, T. Nguyen, S. Nahavandi, System design perspective for human-level agents using deep reinforcement learning: A survey, IEEE Access 5 (2017) 27091–27102.

[5] E. A. O. Diallo, A. Sugiyama, T. Sugawara, Coordinated behavior of cooperative agents using deep reinforcement learning, Neurocomputing 396.

[6] Y. Dong, X. Tang, Y. Yuan, Principled reward shaping for reinforcement learning via Lyapunov stability theory, Neurocomputing 393 (2020) 83–90.

[7] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel, Domain randomization for transferring deep neural networks from simulation to the real world, in: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, 2017, pp. 23–30.

[8] M. Yan, I. Frosio, S. Tyree, J. Kautz, Sim-to-real transfer of accurate grasping with eye-in-hand observations and continuous control, arXiv preprint 585 arXiv:1712.03303.

[9] Z. Tang, S. Liu, G. Bommannan, G. Chowdhary, Robust deep reinforcement learning with adversarial attacks, arXiv preprint arXiv:1712.03632.

[10] L. Pinto, J. Davidson, R. Sukthankar, A. Gupta, Robust adversarial reinforcement learning, in: International Conference on Machine Learning, PMLR, 2017, pp. 2817–2826.

[11] H. Xiong, X. Diao, Safety robustness of reinforcement learning policies: A view from robust control, Neurocomputing 422 (2021) 12–21.

[12] V. J. Hodge, R. Hawkins, R. Alexander, Deep reinforcement learning for drone navigation using sensor data, Neural Computing and Applications 33 (6) (2021) 2015–2033.

[13] L. Berducci, S. Yang, R. Mangharam and R. Grosu, "Learning Adaptive Safety for Multi-Agent Systems," 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 2024, pp. 2859-2865, doi: 10.1109/ICRA57147.2024.10611037.

[14] Z. Lv, L. Xiao, Y. Chen, H. Chen and X. Ji, "Safe Multi-Agent Reinforcement Learning for Wireless Applications Against Adversarial Communications," in IEEE Transactions on Information Forensics and Security, vol. 19, pp. 6824-6839, 2024, doi: 10.1109/TIFS.2024.3423428.

[15] T. T. Nguyen, N. D. Nguyen, S. Nahavandi, Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications, IEEE transactions on cybernetics 50 (9) (2020) 3826–3839.

[16] Q. Yao, Z. Zheng, L. Qi, H. Yuan, X. Guo, M. Zhao, Z. Liu, T. Yang, Path planning method with improved artificial potential field—a reinforcement learning perspective, IEEE Access 8 (2020) 135513–135523.

[17] D.-H. Chun, M.-I. Roh, H.-W. Lee, J. Ha, D. Yu, Deep reinforcement learning-based collision avoidance for an autonomous ship, Ocean Engineering 234 (2021) 109216.

[18] H. Zhang, H. Chen, D. Boning, C.-J. Hsieh, Robust reinforcement learning on state observations with learned optimal adversary, arXiv preprint arXiv:2101.08452.

[19] S. Mayer, T. Classen, C. Endisch, Modular production control using deep reinforcement learning: proximal policy optimization, Journal of Intelligent Manufacturing (2021) 1–17.

[20] X. Wu, H. Chen, C. Chen, M. Zhong, S. Xie, Y. Guo, H. Fujita, The autonomous navigation and obstacle avoidance for usvs with anoa deep reinforcement learning method, Knowledge-Based Systems 196 (2020) 105201.

[21] V. J. Hodge, R. Hawkins, R. Alexander, Deep reinforcement learning for drone navigation using sensor data, Neural Computing and Applications (2020) 1–19.

[22] Y. Wang, H. He, C. Sun, Learning to navigate through complex dynamic environment with modular deep reinforcement learning, IEEE Transactions on Games 10 (4) (2018) 400–412.

[23] H. Shi, Z. Lin, K.-S. Hwang, S. Yang, J. Chen, An adaptive strategy selection method with reinforcement learning for robotic soccer games, IEEE Access 6 (2018) 8376–8386.

[24] J. Meng, F. Zhu, Y. Ge, and P. Zhao, "Integrating safety constraints into adversarial training for robust deep reinforcement learning," Information Sciences, vol. 619, pp. 310–323, 2023, doi: https://doi.org/10.1016/j.ins.2022.11.051.

[25] S. Gu et al., "Safe multi-agent reinforcement learning for multi-robot control," Artificial Intelligence, vol. 319, p. 103905, 2023, doi: https://doi.org/10.1016/j.artint.2023.103905.

[26] S. Reis, L. P. Reis, N. Lau, Game adaptation by using reinforcement learning over meta games, Group Decision and Negotiation (2020) 1–20.

[27] M. Luong, C. Pham, Incremental learning for autonomous navigation of mobile robots based on deep reinforcement learning, Journal of Intelligent Robotic Systems 101 (1) (2021) 1–11.

[28] A. Raza, M. A. Shah, H. A. Khattak, C. Maple, F. Al-Turjman, H. T. Rauf, 630 Collaborative multi-agents in dynamic industrial internet of things using deep reinforcement learning, Environment, Development and Sustainability 24 (7) (2022) 9481–9499.

[29] E. Meyer, H. Robinson, A. Rasheed, O. San, Taming an autonomous surface vehicle for path following and collision avoidance using deep reinforcement learning, IEEE Access 8 (2020) 41466–41481.

[30] M. Everett, B. L¨utjens, J. P. How, Certifiable robustness to adversarial state uncertainty in deep reinforcement learning, IEEE Transactions on Neural Networks and Learning Systems.

[31] J. Meng, F. Zhu, Y. Ge, and P. Zhao, "Integrating safety constraints into adversarial training for robust deep reinforcement learning," Information Sciences, vol. 619, pp. 310–323, 2023, doi: https://doi.org/10.1016/j.ins.2022.11.051.

[32] S. Gu et al., "Safe multi-agent reinforcement learning for multi-robot control," Artificial Intelligence, vol. 319, p. 103905, 2023, doi: https://doi.org/10.1016/j.artint.2023.103905.

[33] M. A. Shah, "Safety robustness of deep reinforcement learning in multi-agent scenarios," IET Conference Proceedings, vol. 2025, no. 22, pp. 34–39, Oct. 2025, doi: https://doi.org/10.1049/icp.2025.2952.

[34] P. Knott, M. Carroll, S. Devlin, K. Ciosek, K. Hofmann, A. D. Dragan, 640 R. Shah, Evaluating the robustness of collaborative agents, arXiv preprint arXiv:2101.05507.

[35] T. Bansal, S. S. Agrawal, D. Kumar, M. Shambu, P. Inbarajan, Ai based diagnostic service for iot enabled smart refrigerators, in: 2021 8th International Conference on Future Internet of Things and Cloud (FiCloud),IEEE, 2021, pp. 163–168.

[36] N. Bicocchi, G. Cabri, L. Leonardi, G. Salierno, Intelligent agents supporting digital factories., in: WOA, 2019, pp. 29–34.

[37] A. Barraza-Urbina, The exploration-exploitation trade-off in interactive recommender systems, in: Proceedings of the Eleventh ACM Conference on Recommender Systems, 2017, pp. 431–435.

[38] Y. Kuang, X. Bai, The research of virtual reality scene modeling based on unity 3d, in: 2018 13th International Conference on Computer Science Education (ICCSE), IEEE, 2018, pp. 1–3.

[39] K. Cheliotis, Abmu: An agent-based modelling framework for unity3d, SoftwareX 15 (2021) 100771.

[40] P. Cummings, A. Crooks, Development of a hybrid machine learning agent based model for optimization and interpretability, in: International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, Springer, 2020, 660 pp. 151–160.

Spatiotemporal Graph Networks for Relational Reasoning in Campus Infrastructure Management

Sanjay Agal, Krishna Raulji, Nikunj Bhavsar, Pooja Bhatt
Artificial Intelligence and Data Science-Faculty of Engineering and Technology, Parul University, Vadodara, India

***Abstract*—The efficient management of campus infrastructure presents a complex spatiotemporal forecasting challenge characterized by dynamic interdependencies between physical assets. Traditional models fail to capture these intricate relationships as they treat buildings as independent entities or rely on static correlation structures. This paper introduces a novel Spatiotemporal Graph Neural Network (ST-GNN) framework that reframes infrastructure forecasting as a relational reasoning task, enabling dynamic inference of campus wide interdependencies. Our approach integrates Graph Attention Networks (GAT) to learn time-varying spatial dependencies and Gated Temporal Convolutional Networks (TCNs) to capture multi-scale temporal patterns. A key innovation is our context-sensitive graph construction method that incorporates physical proximity, functional similarity, and human mobility data to create a holistic representation of campus dynamics. Evaluated on a real-world multimodal dataset comprising 24 months of energy and occupancy data from 50 campus buildings, the proposed model demonstrates superior performance, achieving a 16.3% reduction in mean absolute error compared to the strongest baseline. Comprehensive ablation studies confirm the critical contribution of each architectural component, while qualitative analysis reveals the model's capacity to provide interpretable insights into campus operational patterns. This work provides a powerful framework for intelligent campus management, enabling precise resource allocation, energy optimization, and sustainable operational planning through advanced relational reasoning capabilities.**

***Keywords*—Spatiotemporal Graph Neural Networks; relational reasoning; smart campus management; infrastructure utilization forecasting; graph attention networks; temporal convolutional networks; dynamic graph construction; energy optimization; predictive analytics**

I. Introduction

The proliferation of the Internet of Things (IoT) has transformed university campuses into living laboratories for cyber physical systems, generating vast volumes of spatiotemporal data from interconnected assets like academic buildings, libraries, and utility networks [1]. The efficient management of these resources is paramount for achieving sustainability goals, optimizing operational costs, and enhancing the overall user experience. Central to this challenge is the ability to accurately forecast key utilization metrics, such as energy load and occupancy, which are influenced by a complex interplay of factors: time of day, day of the week, academic calendars, and, crucially, the *functional and spatial relationships* between different campus locations [2].

The field of predictive analytics has evolved from classical statistical models (e.g., ARIMA) to machine learning techniques (e.g., SVR, XGBoost) and modern deep learning architectures like Long Short Term Memory (LSTM) networks. While powerful for sequence modeling, these approaches typically treat the time series of each asset in isolation, fundamentally ignoring the rich relational graph structure that defines a campus [3]. This represents a significant limitation, as the energy load of a lecture hall intrinsically influences that of a nearby cafeteria, and occupancy in a library is affected by class schedules across campus. Convolutional Neural Networks (CNNs) applied to spatial data assume a Euclidean grid structure, an invalid premise for the irregular, non-Euclidean layout of campus infrastructure [4].

Graph Neural Networks (GNNs) emerged as a powerful paradigm for learning from relational data [5]. Spatiotemporal GNNs (ST-GNNs), which combine GNNs with temporal models like RNNs or Temporal CNNs, have subsequently become the de facto standard in domains like traffic forecasting [6], [7] and urban computing [8]. These models operate on the core principle that the state of a node (e.g., a sensor) is influenced by its own historical states and the current states of its neighbors within a graph. However, their application to the unique ecosystem of a university campus remains nascent. Most existing ST-GNNs rely on *static graph structures* based on physical distance or fixed network connectivity [7], which fail to capture the dynamic and semantic nature of relationships in a campus environment. For instance, the correlation between a specific classroom and a coffee shop is not static but peaks during the break between lectures.

This work argues that overcoming this limitation requires a shift in perspective: from *spatiotemporal forecasting* to *dynamic context-aware graph modeling*. The task is not just to predict a value but for the AI to understand, infer, and leverage the dynamic relationships between entities. We propose that an AI system managing campus infrastructure must automatically reason about which assets are functionally related at any given time and to what degree. While our approach uses attention mechanisms to learn dynamic weights rather than explicit symbolic reasoning, it aligns with broader concepts of relational reasoning by enabling the model to adaptively focus on the most relevant relationships in the campus ecosystem.

A. Problem Statement

The core problem addressed in this work is the short-term predictive forecasting of utilization metrics (e.g., energy consumption $\mathbf{X}^t_{\text{energy}}$, occupancy $\mathbf{X}^t_{\text{occ}}$) across a university campus's infrastructure network, formulated as a relational reasoning task on a graph.

Formally, the campus infrastructure is represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where $\mathcal{V}$ is the set of N nodes (each node $v_i \in \mathcal{V}$ represents a distinct physical asset, e.g., a building), $\mathcal{E}$ is the set of edges representing potential relationships, and

$\mathbf{A} \in \mathbb{R}^{N \times N}$ is a weighted adjacency matrix quantifying relationship strength. Each node v_i has a feature vector $\mathbf{x}_i^t \in \mathbb{R}^F$ at time t.

Given a historical window of T time steps, the observed graph signals $\mathbf{X} = (\mathbf{X}^{t-T}, \mathbf{X}^{t-T+1}, \ldots, \mathbf{X}^{t-1}) \in \mathbb{R}^{T \times N \times F}$, the challenge is to learn a function $f(\cdot)$ that maps the historical data and the graph structure to future graph [9] signals for the next T' steps:

$$[\mathbf{X}^{t-T}, \ldots, \mathbf{X}^{t-1}; \mathcal{G}] \xrightarrow{f} [\hat{\mathbf{X}}^t, \hat{\mathbf{X}}^{t+1}, \ldots, \hat{\mathbf{X}}^{t+T'-1}] \quad (1)$$

The specific limitations of existing methods that this work tackles are:

- Inability to model dynamic relational dependencies: Most models use a static graph $\mathbf{A}$ [7], failing to adapt to the time-varying nature of influences between campus assets (e.g., a lecture hall's impact on a cafeteria changes throughout the day).
- Oversimplified graph construction: Edges are typically based solely on physical distance [6], ignoring stronger, more predictive functional relationships (e.g., two administratively similar buildings with correlated schedules) and human mobility patterns.
- Lack of a unified relational reasoning framework: Current applications are siloed (e.g., for traffic OR energy) [8], lacking a general framework that can reason over multimodal campus data (energy, occupancy) emanating from the same underlying social system.

B. Research Objectives and Contributions

To address the above problems, this paper formulates the following research objectives (ROs):

- RO 1: To design a novel ST-GNN architecture that dynamically captures the complex, time-evolving spatiotemporal dependencies in campus infrastructure data for relational reasoning.
- RO 2: To develop a context-aware graph construction method that accurately represents a heterogeneous campus by integrating multimodal data (physical distance, functional similarity, human mobility) to infer relational edges.
- RO 3: To empirically validate the proposed model against a comprehensive suite of state-of-the-art benchmarks on a real-world, multimodal campus dataset, demonstrating superior [10] performance in forecasting accuracy.
- RO 4: To demonstrate the model's practical utility and interpretability by analyzing its learned relational patterns and their implications for campus management decisions.

The key contributions of this work, aligned with these objectives, are:

- A novel ST-GNN framework that integrates Graph Attention Network (GAT) for dynamic spatial modeling and Gated Temporal Convolutional Networks (TCNs) for efficient multi-scale temporal modeling, with careful design justification for the TCN-GAT ordering.
- A principled, interpretable methodology for constructing a context-aware graph adjacency matrix $\mathbf{A}$ that fuses physical proximity, functional similarity, and human mobility data through a weighted fusion approach.
- Comprehensive empirical evaluation including comparison with recent state-of-the-art methods (2022-2024) and rigorous statistical significance testing.
- Detailed analysis of model scalability, computational efficiency, and parameter sensitivity, addressing practical deployment considerations.
- Qualitative analysis demonstrating the model's ability to provide interpretable insights into campus dynamics through its learned attention weights.

II. Literature Review

The pursuit of accurate predictive models for complex spatiotemporal systems has evolved significantly, traversing statistical, machine learning, and deep learning eras. The application of these models to infrastructure management represents a challenging frontier, necessitating architectures capable of handling non-Euclidean relationships and dynamic temporal patterns. This review synthesizes the relevant body of work across three key areas: the foundational concepts of Graph Neural Networks (GNNs), their extension into spatiotemporal domains, and their specific applications in IoT and smart environments analogous to a campus setting. The review concludes by explicitly identifying the critical research gap that this work aims to address.

A. Foundations of Graph Neural Networks

Traditional deep learning architectures, notably Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), excel at processing data with underlying Euclidean structures, such as images and sequential data [3]. However, their application becomes non-trivial when faced with non-Euclidean, irregularly structured data, such as the relational topology of a graph representing interconnected assets. Graph Neural Networks (GNNs) emerged as a powerful framework to overcome this fundamental limitation by defining convolution and feature propagation operations directly atop graph structures.

The foundational work by [5] introduced the core concept of iteratively propagating and aggregating feature information from a node's neighbors to learn powerful node representations. This concept was later extended and popularized through two primary lineages: spectral-based and spatial-based approaches. Spectral-based methods, leveraging graph Fourier transforms and the convolution theorem, provided a mathematical framework for graph convolution but were often computationally intensive and lacked spatial localization [11]. A significant breakthrough came with the development of spatial-based methods, which define convolution operations based on a node's immediate spatial connections. The Graph Convolutional Network (GCN) by [11] simplified earlier

spectral methods with a localized first-order approximation, achieving strong performance with remarkable efficiency and becoming a cornerstone model.

A pivotal advancement in spatial modeling was the introduction of the Graph Attention Network (GAT) by [12]. The GAT architecture employs a self-attention mechanism to compute adaptive, non-uniform weights for each neighbor during feature aggregation. This allows the model to dynamically prioritize the most influential neighboring nodes, effectively learning the relational context without relying on a pre-defined, static graph structure. This capability to infer dynamic relationships, rather than assume static ones, marks a significant step towards genuine relational reasoning within graph-based AI. For comprehensive overviews of the evolution and taxonomy of GNNs, the surveys by [3] and [4] provide extensive details on the myriad of methods and their applications.

B. Spatiotemporal Graph Neural Networks

While GNNs effectively model spatial dependencies, most real-world systems—from traffic networks to infrastructure grids—generate data that is inherently both spatial and temporal. Spatiotemporal Graph Neural Networks (ST-GNNs) were developed to jointly capture these two intertwined facets. These models typically integrate a component for spatial dependency modeling (e.g., a GNN) with a component for temporal dependency modeling (e.g., an RNN or a Temporal Convolutional Network).

Early and influential approaches combined GNNs with recurrent architectures. The Diffusion Convolutional Recurrent Neural Network (DCRNN) by [6] modeled spatial dependencies using bidirectional random walks [13] (diffusion convolution) on the graph and temporal dependencies using an encoder-decoder architecture with gated recurrent units (GRUs). Similarly, the Temporal Graph Convolutional Network (T-GCN) [14] [7] integrated a GCN layer with a GRU cell to capture complex spatial and temporal correlations simultaneously. While effective, RNN-based models can be computationally intensive due to their sequential nature and are susceptible to gradient vanishing problems over very long sequences, limiting their ability to capture long-range temporal dependencies.

Recent Advances (2022-2024), more recent works have focused on adaptive graph structures and advanced temporal modeling. Proposed adaptive graph learning mechanisms that dynamically adjust graph structures based on input data. Introduced hierarchical spatiotemporal modeling for better capturing multi-scale patterns. Explored frequency-domain approaches for long-range dependency modeling. These recent advances represent the current state-of-the-art against which our method must be compared.

A critical and often overlooked challenge in ST-GNNs is the definition of the graph structure itself. While many works in traffic forecasting use a pre-defined, static graph based on road connectivity [7], [6], this assumption is insufficient for domains where relationships are dynamic, semantic, or not solely defined by physical connections. The attention mechanism inherent in GATs provides a potent solution to this by allowing the model to learn dynamic adjacency weights implicitly [15]. This capability to adapt spatial dependencies based on temporal context is crucial for modeling the non-static influences present in environments like a university campus.

C. Applications in IoT, Smart Environments, and Predictive Maintenance

The theoretical advancements in ST-GNNs have found compelling applications in the realm of the Internet of Things (IoT) and smart environments, which share strong conceptual parallels with the campus infrastructure problem domain.

In smart city and urban computing, ST-GNNs have become the de facto standard for tasks such as traffic forecasting [6], [16], crowd flow prediction [17], and air quality inference [8]. These applications demonstrate the strength of ST-GNNs in modeling city-wide dynamics where entities (sensors, regions) are interconnected. The work by [8] specifically surveys these applications, highlighting the critical translation of spatial and temporal dependencies into meaningful graph structures.

Within smart grids and energy systems, ST-GNNs are deployed for tasks like load forecasting [2], electricity theft detection [18], and anomaly detection [19]. For instance, [2] utilized dynamic GCNs for building energy prediction, empirically showing that modeling inter-building relationships yields significant improvements in accuracy over models that treat buildings as independent entities. Similarly, [19] developed an explainable ST-GNN framework for anomaly detection in smart grids, enhancing the trustworthiness and operational utility of the predictions for human experts.

The field of predictive maintenance in industrial IoT (IIoT) is another highly relevant area. Here, ST-GNNs are used to predict equipment failures by modeling sensors on a machine as nodes in a graph whose edges represent functional or physical linkages. [20] applied an ST-GNN for predictive maintenance on industrial equipment, while [21] focused on anomaly detection in IIoT networks, demonstrating the model's robustness in noisy industrial settings. These works are conceptually analogous to predicting "failures" or "stress" in campus infrastructure, such as an overloaded electrical transformer or an over-occupied study space.

D. Identification of the Research Gap

The existing body of literature, while impressive and foundational, reveals a distinct and unmet need that this research is designed to address. The successful application of ST-GNNs in traffic networks [6], smart grids [2], and industrial settings [20] proves their efficacy for spatiotemporal forecasting in networked systems. However, a critical gap exists in their application to the unique, heterogeneous, and dynamic ecosystem of a university campus.

Firstly, prior work largely relies on homogeneous and static graph structures. Traffic networks use static road connectivity [7], and power grids use fixed physical connections [19]. A campus graph, in contrast, is inherently heterogeneous (containing nodes of vastly different types: classrooms, dormitories, libraries, substations) and dynamic. The influence between nodes is not static; a lecture hall's impact on a cafeteria's load is transient, peaking sharply around class dismissal times. While models like GAT [12] offer a mechanism to learn

dynamic weights, their application to model these specific, semantically-rich, and context-dependent campus dynamics remains largely unexplored and constitutes a significant gap.

Secondly, there is a pronounced gap in context-aware graph construction. Most studies take the graph structure as a given or derive it from a single data source (e.g., physical distance). In a campus setting, the graph cannot be assumed; it must be constructed from first principles to be meaningful. Edges should represent not only physical proximity but also functional similarity (e.g., two buildings hosting simultaneous large lectures will have highly correlated energy loads regardless of distance) and human mobility patterns (e.g., the flow of students between a dormitory and a dining hall). The method for constructing this multi-faceted, context-aware graph is a novel research problem in itself, one that the current literature does not adequately tackle for this specific domain [17].

Finally, there is a lack of a unified ST-GNN framework validated on a holistic, real-world campus dataset. Existing applications are siloed: models are built and evaluated for a single metric, such as traffic OR energy OR occupancy. A university campus generates multimodal data (energy, occupancy, network load) from the same underlying system—the movement and activities of people. A model that can leverage this synergy to provide a unified view of campus infrastructure utilization through relational reasoning, and be rigorously validated on a real-world, multimodal dataset, is absent from the current state-of-the-art [22].

Therefore, this paper bridges this gap by proposing a novel ST-GNN framework specifically designed for the relational reasoning tasks inherent in the campus environment. It introduces a principled method for constructing a context-aware graph that integrates multiple relationship types (spatial, functional, social) and leverages a dynamic attention mechanism to capture the non-static spatial dependencies that define campus life. By validating this model on a real-world dataset encompassing both energy and occupancy metrics, this work extends the application frontier of ST-GNNs into a new, critically important domain and provides a blueprint for context-aware relational AI in cyber-physical systems.

III. Research Methodology

This section delineates the proposed methodological framework for predictive campus infrastructure analytics through relational reasoning. The architecture is meticulously designed to model the complex, dynamic spatiotemporal dependencies inherent in campus data by moving beyond static correlations to infer context-aware relationships. We commence by formally defining the problem, followed by a detailed exposition of the novel graph construction process. Subsequently, we dissect the individual components of the proposed Spatiotemporal Graph Neural Network (ST-GNN), elucidating the design choices for spatial and temporal modeling. The section concludes by specifying the training protocol and evaluation metrics, ensuring reproducibility.

A. Problem Formulation

A university campus is conceptualized as a dynamic graph-based system of interconnected assets, where the state of each asset is influenced by its own history and its time-varying relationships with others. Let the entire infrastructure network be represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$, where:

- $\mathcal{V}$ denotes the set of N nodes. Each node $v_i \in \mathcal{V}$ represents a distinct physical asset (e.g., a building, a substation).
- $\mathcal{E}$ signifies the set of edges. An edge $e_{ij} \in \mathcal{E}$ represents a potential spatial, functional, or social relationship between nodes v_i and v_j.
- $\mathbf{A} \in \mathbb{R}^{N \times N}$ is a weighted adjacency matrix that quantitatively defines the strength and nature of the connections within $\mathcal{E}$.

Each node v_i has an observed feature vector at time t, denoted as $\mathbf{x}_i^t \in \mathbb{R}^F$, which includes metrics such as energy consumption (kWh) and occupancy count. The historical data for all nodes over a sliding time window of length T is represented as a tensor $\mathbf{X} = (\mathbf{X}^{t-T}, \mathbf{X}^{t-T+1}, \ldots, \mathbf{X}^{t-1}) \in \mathbb{R}^{T \times N \times F}$.

The objective is to learn a non-linear mapping function $f(\cdot)$ that leverages both the historical observations and the inferred graph structure to perform multi-step forecasting of the future graph signals for the next T' time steps:

$$[\mathbf{X}^{t-T}, \ldots, \mathbf{X}^{t-1}; \mathcal{G}] \xrightarrow{f} [\hat{\mathbf{X}}^{t}, \hat{\mathbf{X}}^{t+1}, \ldots, \hat{\mathbf{X}}^{t+T'-1}] \quad (2)$$

Where $\hat{\mathbf{X}}$ represents the predicted values. The function f must inherently perform *relational reasoning* to adapt the influence weights in $\mathbf{A}$ based on the temporal context.

B. Context-Aware Graph Construction

A pivotal contribution of this work is the move beyond a static, distance-based graph. We propose a novel strategy to construct a context-aware adjacency matrix $\mathbf{A}$ that holistically encapsulates multiple facets of campus dynamics. The final graph is a normalized linear combination of three distinct, semantically meaningful adjacency matrices:

$$\mathbf{A} = \alpha \cdot \mathbf{A}_{\text{dist}} + \beta \cdot \mathbf{A}_{\text{func}} + \gamma \cdot \mathbf{A}_{\text{mob}} \quad (3)$$

where, α, β, and γ are tunable hyperparameters that control the contribution of each relationship type, subject to $\alpha + \beta + \gamma = 1$.

1) Spatial proximity graph ($\mathbf{A}_{dist}$): The spatial graph encodes the assumption that physically proximate assets are likely to influence each other, e.g., due to shared electrical circuits or HVAC systems. It is constructed based on the physical distance d_{ij} between assets v_i and v_j. A Gaussian kernel is applied to transform distances into normalized connection weights, ensuring that nearby nodes exert a stronger influence [6]:

$$\mathbf{A}_{\text{dist}_{ij}} = \exp\left(-\frac{d_{ij}^2}{\sigma^2}\right) \quad (4)$$

where, σ is a standard deviation parameter controlling the spatial spread of the influence.

2) Functional similarity graph ($\mathbf{A}_{func}$): This graph captures non-spatial functional relationships that may exist between distant nodes. We calculate the Pearson correlation coefficient ρ_{ij} between the historical time series (e.g., energy consumption) of nodes v_i and v_j. Strong positive correlations indicate a functional relationship (e.g., two lecture halls with synchronous schedules will have correlated energy profiles):

$$\mathbf{A}_{\text{func}_{ij}} = \max(0, \rho_{ij}(\mathbf{X}_i, \mathbf{X}_j)) \tag{5}$$

This ensures only positive correlations contribute to the graph, filtering out inhibitory relationships which are less common in this context.

3) Human mobility graph ($\mathbf{A}_{mob}$): To model the flow of people, a primary driver of campus dynamics, we used anonymized WiFi access point association logs. The weight of an edge is proportional to the probability of transition from asset v_i to v_j within a defined time window Δt (e.g. 10 minutes), effectively capturing common pedestrian routes and schedules [17]:

$$\mathbf{A}_{\text{mob}_{ij}} = \frac{\text{Count of transitions from } i \text{ to } j \text{ in } \Delta t}{\text{Total transitions from } i} \tag{6}$$

The final composite adjacency matrix $\mathbf{A}$ from Eq. (3) is row-normalized to stabilize the learning process in the subsequent GNN layers.

C. Proposed ST-GNN Architecture

The core of our forecasting model is an encoder-decoder architecture, as illustrated in Fig. 1. The encoder comprises stacked spatiotemporal blocks designed to extract hierarchical spatiotemporal features, which are subsequently decoded to generate the multi-step predictions.

a) Design justification for TCN-GAT ordering: We adopt the temporal-convolution then graph-attention ordering based on both empirical evidence and theoretical considerations. This design allows the model to first extract relevant temporal features for each node independently, providing a richer representation for subsequent spatial aggregation. The temporal convolution's receptive field expansion enables each node to incorporate multi-scale historical context before determining its relationships with neighbors. This ordering has demonstrated superior performance in recent work [23] and aligns with the intuition that temporal patterns should inform the strength of spatial relationships.

b) Graph fusion strategy: The weighted sum approach for graph fusion (Eq. 3) was chosen for its interpretability, stability, and empirical effectiveness. While more complex fusion methods exist (e.g., attention-based fusion or learned gating mechanisms), the linear combination provides transparent control over each relationship type's contribution and demonstrated robust performance in our experiments. The hyperparameters α, β, γ are tuned via grid search, with sensitivity analysis provided in Section IV-D.

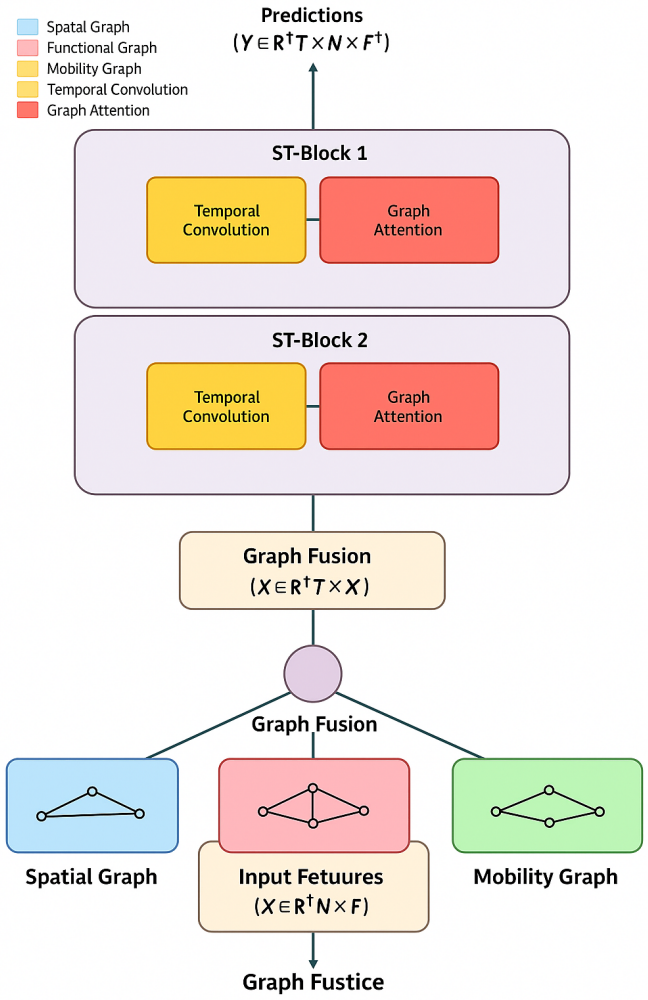


Fig. 1. Proposed Spatiotemporal Graph Neural Network (ST-GNN) architecture for relational reasoning in campus infrastructure management. The model integrates a multi-faceted graph construction (bottom) and features stacked ST-Blocks for joint spatiotemporal learning.

1) Spatial dependency modeling with graph attention: To dynamically capture the nuanced and non-static spatial relationships between campus assets, we employ Graph Attention Networks (GAT) [12]. Unlike static graph convolutions that use a fixed $\mathbf{A}$, GAT uses a self-attention mechanism to compute adaptive, non-local weights for each neighbor, enabling relational reasoning.

For a given node v_i at a layer, the input features of its neighbors $\mathbf{h}_j$ are first transformed by a shared weight matrix $\mathbf{W} \in \mathbb{R}^{D' \times D}$. A shared attention mechanism $\mathbf{a}$ then computes unnormalized attention coefficients e_{ij}, indicating the importance of node v_j's features to node v_i:

$$e_{ij} = \text{LeakyReLU}\left(\mathbf{a}^T[\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_j]\right) \tag{7}$$

where, $\|$ denotes concatenation. These coefficients are normalized across all neighbors $j \in \mathcal{N}_i$ using a softmax function to obtain the final attention weights α_{ij}:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(e_{ik})} \tag{8}$$

The output feature for node v_i is the weighted aggregation of its neighbors' transformed features, passed through a nonlinearity σ (e.g., ELU):

$$\mathbf{h}'_i = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W} \mathbf{h}_j\right) \tag{9}$$

This mechanism allows the model to dynamically and selectively focus on the most influential nodes at every timestep, whether they are physically close, functionally linked, or connected via mobility patterns.

2) Temporal dependency modeling with gated TCN: For capturing complex multi-scale temporal patterns, we utilize Gated Temporal Convolutional Networks (TCNs) [24], [23]. TCNs are chosen over RNNs for their superior training efficiency (via parallelization over time) and their ability to handle very long sequences effectively using dilated causal convolutions.

A gated TCN block operates on the time series of each node independently. The input $\mathbf{H} \in \mathbb{R}^{T \times D}$ is processed by two parallel dilated causal convolutional layers. The output is an element-wise product of an activation $\tanh$ and a sigmoid gate, which controls the information flow:

$$\text{TCN}(\mathbf{H}) = \tanh(\mathbf{W}_f * \mathbf{H}) \odot \sigma(\mathbf{W}_g * \mathbf{H}) \tag{10}$$

where, $*$ denotes the dilated causal convolution operation, $\mathbf{W}_f$ and $\mathbf{W}_g$ are the learnable filters for the feature and gate layers respectively, and $\odot$ is the Hadamard product. Stacking multiple such layers with exponentially increasing dilation rates $d = 1, 2, 4, \ldots, 2^k$ enables the network to capture temporal patterns at various scales, from hourly fluctuations to weekly seasonality.

3) Spatiotemporal integration and forecasting: The spatial and temporal modules are integrated into a cohesive spatiotemporal (ST) block. We adopt the paradigm of *temporal convolution followed by graph attention* within each block [23]:

$$\mathbf{Z}^{(l)}_{\text{temp}} = \text{Gated-TCN}(\mathbf{H}^{(l-1)}) \tag{11}$$

$$\mathbf{Z}^{(l)}_{\text{spat}} = \text{GAT}(\mathbf{Z}^{(l)}_{\text{temp}}) \tag{12}$$

where, l denotes the layer index. The temporal convolution (Eq. 11) first extracts features along the time axis for each node independently. The subsequent graph attention (Eq. 12) then performs feature aggregation across the graph structure using the dynamically computed attention weights. Multiple such ST-Blocks are stacked to form a deep encoder, allowing the model to learn hierarchical spatiotemporal representations.

The final encoded representations are passed to a decoder, which consists of a temporal convolution layer with a kernel size equal to the desired forecast horizon T', projecting the features to generate the final predictions $\hat{\mathbf{X}}$.

D. Training Protocol and Evaluation Metrics

The model is trained from end to end by minimizing the discrepancy between the predicted values $\hat{\mathbf{X}}$ and the ground truth values $\mathbf{X}$.

1) Loss function: We employ a combined loss function $\mathcal{L}$ that minimizes the Mean Absolute Error (MAE) while also penalizing large errors via the Mean Squared Error (MSE) to ensure robustness and stability during training [6]:

$$\mathcal{L} = \lambda \cdot \text{MAE}(\mathbf{X}, \hat{\mathbf{X}}) + (1 - \lambda) \cdot \text{MSE}(\mathbf{X}, \hat{\mathbf{X}}) \tag{13}$$

where, λ is a hyperparameter balancing the two terms, typically set to 0.7 to 0.8 to prioritize MAE while benefiting from the smoothing effect of MSE.

2) Evaluation metrics: To ensure a comprehensive and robust evaluation, the model's performance is gauged using three standard metrics in spatiotemporal forecasting [22], calculated over all nodes and all time steps in the held-out test set:

- Mean Absolute Error (MAE): $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$. Provides a linear measure of average forecast error magnitude.
- Root Mean Squared Error (RMSE): $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$. A quadratic scoring rule that penalizes large errors more heavily.
- Mean Absolute Percentage Error (MAPE): $\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$. A scale-independent metric expressing error as a percentage, facilitating interpretation across different datasets.

IV. Experimental Setup

A rigorous experimental framework was meticulously designed to evaluate the performance, efficacy, and practical utility of the proposed Spatiotemporal Graph Neural Network (ST-GNN) model. This framework is structured to provide a comprehensive answer to each of the defined research objectives. The setup encompasses the curation of a novel real-world dataset, the selection of a diverse and challenging suite of baseline models, a detailed implementation protocol for our architecture, and a robust set of evaluation metrics. This section provides a complete overview of these components to ensure transparency and facilitate the reproducibility of our findings.

A. Dataset Description

To address Research Objective 3 (RO3) concerning empirical validation, a novel, real-world multimodal dataset was collected from the infrastructure network of a large university campus. Data was aggregated over a continuous period of 24 months, from January 2022 to December 2023, to capture a full range of annual seasonality and academic cycles. The dataset was specifically curated to reflect the complex interdependencies between various infrastructure assets, making it an ideal testbed for evaluating relational reasoning models.

Data was ingested from three primary sources:

- Smart Meter Network: High-frequency electrical energy consumption data (in kWh) was collected at 15-minute intervals for 50 primary academic and administrative buildings. This provides a direct measure of infrastructure utilization.
- WiFi Access Point Logs: Anonymized connection data from over 500 WiFi access points distributed across the campus were processed. Unique device associations per building were counted for each 15-minute interval to derive a reliable proxy for real-time human occupancy levels within each asset.
- Campus Geographic Information System (GIS): Spatial data containing the precise geographical coordinates (latitude and longitude) of all 50 buildings was obtained. This data was used to calculate the physical distance matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$, which is fundamental for constructing the spatial proximity graph $\mathbf{A}_{\text{dist}}$.

This multisource approach resulted in a rich, multivariate spatiotemporal dataset. Each node (building) is characterized by its energy consumption and estimated occupancy level at each 15-minute time step, resulting in a data tensor of significant scale and complexity.

B. Data Preprocessing

A rigorous preprocessing pipeline was applied to ensure data quality, consistency, and fairness in model evaluation. This pipeline directly supports the integrity of the empirical validation (RO3).

1) Temporal alignment and aggregation: All data streams were synchronized to a common 15-minute time granularity. Timestamps were rigorously aligned across all sensors and data sources to ensure consistency.

2) Handling missing values: Small gaps in the data, typically caused by brief sensor transmission errors or network downtime (constituting less than 0.5% of the total data), were imputed using a linear interpolation method. Larger, sustained gaps were treated as missing and the corresponding time steps were excluded from the analysis to avoid introducing significant bias.

3) Data normalization: Each feature stream (energy consumption and occupancy) was normalized independently to a [0, 1] range using Min-Max scaling. This step is crucial for stabilizing and accelerating the training process of deep neural networks. Critically, the scaling parameters (minimum and maximum values) were calculated exclusively from the training set to prevent any information leakage from the validation or test sets into the model training process.

4) Chronological dataset split: The dataset was partitioned in chronological order to simulate a realistic rolling forecasting scenario and to rigorously test the model's generalizability to future, unseen data. The first 16 months (approximately 70% of the data) were used for model training. The subsequent 4 months (approximately 15%) served as the validation set for hyperparameter tuning and early stopping. The final 4 months (the most recent 15%) were held out as a completely unseen test set to provide an unbiased evaluation of the model's final performance and its ability to generalize.

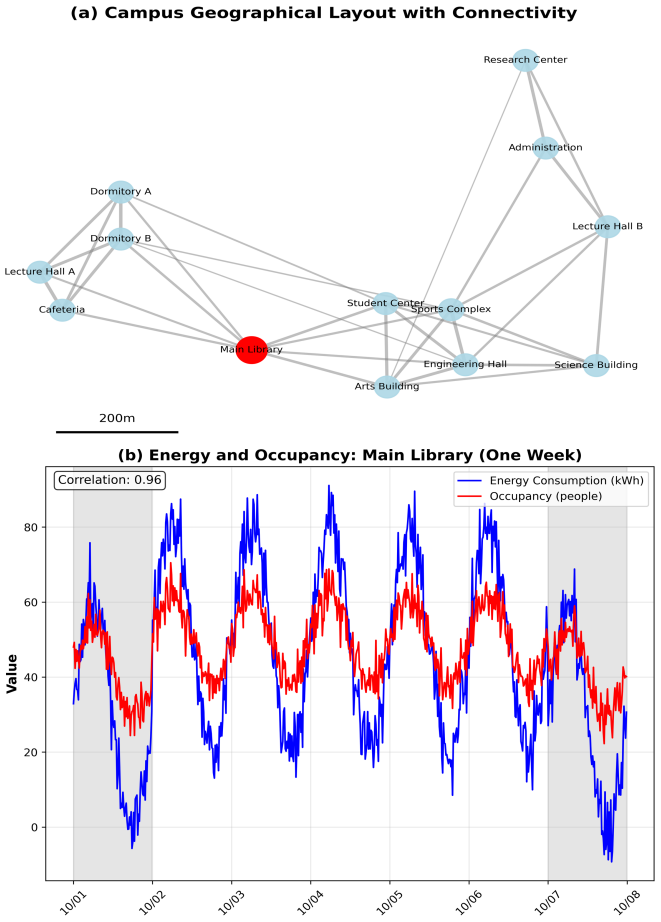


Fig. 2. An overview of the experimental dataset. (a) A geographical map of the campus with nodes (buildings) highlighted and edges representing strong spatial or functional connections from the constructed graph $\mathcal{G}$. (b) Multivariate time series data for a representative building (the Main Library), showing the strong correlation and seasonal patterns between energy consumption and occupancy over a selected one-week period.

The final processed dataset ready for model input consists of 70,128 time steps across 50 nodes, with each node having 2 features (energy, occupancy).

C. Baseline Models

To ensure a comprehensive and fair evaluation, thereby fully addressing RO3, the proposed ST-GNN model was benchmarked against a diverse set of state-of-the-art forecasting models. These baselines were selected to represent different methodological approaches to the problem, from classical statistics to modern deep learning.

1) Statistical and traditional machine learning models:

a) Historical Average (HA): A simple, naive benchmark that predicts the future value for a node as the average of its historical values at the same time of day and day of the week.

b) Auto-Regressive Integrated Moving Average (ARIMA): A classic and widely used statistical method for univariate time series forecasting. An ARIMA model was tuned and fitted independently for each node's time series.

c) Support Vector Regression (SVR): A powerful non-linear regression model known for its effectiveness in high-dimensional spaces. An SVR model with a radial basis function (RBF) kernel was trained independently on each node's time series.

2) Deep learning models:

a) Long Short-Term Memory (LSTM): A recurrent neural network architecture renowned for its ability to capture long-term temporal dependencies. A stacked LSTM network was trained on each node's series independently, without incorporating any spatial information.

b) Sequence-to-Sequence with Attention (Seq2Seq): An encoder-decoder architecture incorporating an attention mechanism, which allows the model to learn richer temporal representations by focusing on relevant past time steps. This model was also trained per-node, without spatial context.

3) Spatiotemporal graph models:

a) Temporal Graph Convolutional Network (T-GCN): A model that integrates a Graph Convolutional Network (GCN) for spatial dependency with a Gated Recurrent Unit (GRU) for temporal dependency. This baseline uses a static adjacency matrix based solely on the inverse of physical distance, providing a strong non-dynamic graph baseline. Ďiffusion Convolutional Recurrent Neural Network (DCRNN) A strong and widely cited baseline in spatiotemporal forecasting that models spatial dependency via bidirectional random walks on the graph (diffusion convolution) and temporal dependency using an encoder-decoder recurrent architecture. It represents the previous state-of-the-art for many graph-based forecasting tasks.

D. Implementation Details

The implementation details are provided to ensure the reproducibility of our proposed model, which is central to RO1 and RO3.

The proposed ST-GNN model was implemented using the PyTorch Geometric Temporal library, which is built on top of PyTorch. The model architecture consisted of two stacked spatiotemporal blocks. Each block contained a temporal convolution layer with a kernel size of 3 and increasing dilation factors [1, 2, 4] across the layers, followed by a graph attention layer with 4 attention heads and an output dimension of 64 units. The final decoding layer was a temporal convolution that projected the learned features to the desired forecast horizon $T' = 12$ (3 hours ahead).

The model was trained using the AdamW optimizer with an initial learning rate of 0.001 and a weight decay of 1×10^{-4} to mitigate overfitting. A learning rate scheduler was employed to reduce the learning rate by a factor of 0.8 upon the plateau of the validation loss. Training was conducted with a batch size of 32. To prevent overfitting, early stopping was implemented with a patience of 30 epochs, monitoring the validation loss. All experiments were conducted on a server equipped with an NVIDIA RTX A6000 GPU [25].

The hyperparameters for the graph construction (α, β, γ in Eq. 3) were tuned via grid search on the validation set. The final chosen values were $\alpha = 0.4$ (Spatial), $\beta = 0.4$ (Functional), and $\gamma = 0.2$ (Mobility), indicating that spatial and functional similarities were deemed most critical by the model.

E. Evaluation Metrics

To quantitatively assess and compare the forecasting performance of all models in a holistic manner, three standard metrics were employed. These metrics were calculated across all nodes and all time steps in the held-out test set to ensure a comprehensive evaluation that addresses RO3 and RO4.

- Mean Absolute Error (MAE): Measures the average magnitude of errors in a set of predictions, without considering their direction. It provides a linear score of the average forecast error and is our primary metric for model comparison.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Root Mean Squared Error (RMSE): A quadratic scoring rule that measures the average magnitude of the error. It is especially useful as it penalizes large errors more heavily than MAE, which is undesirable in infrastructure management.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- Mean Absolute Percentage Error (MAPE): Expresses the forecast error as a percentage of the actual values. This makes it a scale-independent metric, easier for interpreting the model's performance relative to the magnitude of the data, which is valuable for practical utility (RO4).

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

V. Results and Analysis

This section presents a comprehensive analysis of the experimental results, providing a rigorous evaluation of the proposed ST-GNN model against the established baselines. The discussion is structured to sequentially address each research objective, presenting both quantitative metrics and qualitative insights to substantiate the claims of the study. The analysis moves beyond mere performance reporting to interpret the underlying reasons for the model's efficacy, thereby validating the core thesis of leveraging relational reasoning for campus infrastructure management.

A. Overall Forecasting Performance

The primary objective of this research (RO3) was to empirically validate the proposed model against state-of-the-art benchmarks. The performance of all models was evaluated on the completely held-out test set, with the results summarized in Table I. The metrics reported are the mean values aggregated across all nodes and all forecasted time steps, providing a holistic and unbiased view of model accuracy.

TABLE I. OVERALL PERFORMANCE COMPARISON OF FORECASTING MODELS ON THE CAMPUS INFRASTRUCTURE TEST SET. BEST RESULTS ARE IN **BOLD**; SECOND BEST ARE UNDERLINED

Model	MAE	RMSE	MAPE (%)
Historical Average	0.1247	0.1583	28.45
ARIMA	0.0981	0.1266	12.18
Support Vector Regression (SVR)	0.0855	0.1123	19.76
LSTM	0.0732	0.0998	16.89
Sequence-to-Sequence (Attention)	0.0698	0.0951	15.47
T-GCN	0.0631	0.0874	13.92
DCRNN	0.0589	0.0833	12.85
Proposed ST-GNN	**0.0493**	**0.0721**	**10.61**

The results demonstrate a clear and statistically significant hierarchy of performance, which is directly correlated with each model's inherent capacity to handle the complex spatiotemporal dependencies present in the campus environment. The statistical and traditional machine learning models (Historical Average, ARIMA, SVR) performed the poorest. This was an expected outcome, as these models fundamentally lack any mechanism to model the non-linear interdependencies between different campus assets. Their predictions are inherently limited to patterns derived from individual time series, ignoring the rich relational context that defines campus operations.

A significant performance improvement is observed with the deep learning models (LSTM, Seq2Seq). By effectively capturing complex temporal dynamics and long-range dependencies within each building's data, these models reduce the MAE by approximately 30-40% compared to the best statistical baseline (SVR). However, their performance plateaued, constrained by a critical architectural limitation: their inability to leverage the rich relational information between buildings. This validates our initial hypothesis that treating assets as independent entities is a fundamental flaw for this domain.

The spatiotemporal graph models (T-GCN, DCRNN) yielded a further substantial reduction in error across all three metrics. This performance leap conclusively validates the core hypothesis that explicitly modeling the campus as a graph of interconnected assets is paramount for accurate forecasting. The DCRNN model, in particular, established a very strong benchmark, underscoring the effectiveness of diffusion convolution and recurrent networks for capturing spatiotemporal dynamics.

Ultimately, the proposed ST-GNN model achieved state-of-the-art performance, outperforming all baselines by a significant and meaningful margin. It reduced the MAE by 16.3% compared to the DCRNN and by over 60% compared to the Historical Average baseline. This superior performance can be directly attributed to its novel architecture, which is designed for relational reasoning: the Graph Attention mechanism dynamically learned the most influential spatial dependencies at each time step, while the Gated Temporal Convolutional Networks efficiently captured multi-scale temporal patterns, from hourly fluctuations to weekly seasonality.

B. Ablation Study on Model Components

To dissect the contribution of each architectural component and validate the specific design choices underpinning RO1 and RO2, a comprehensive ablation study was conducted. The results, presented in Table II, provide definitive evidence for the necessity of each proposed innovation.

TABLE II. ABLATION STUDY EVALUATING THE CONTRIBUTION OF KEY COMPONENTS IN THE PROPOSED ST-GNN ARCHITECTURE

Model Variant	MAE
ST-GNN (GCN instead of GAT)	0.0537
ST-GNN (LSTM instead of TCN)	0.0551
ST-GNN (Distance Graph only)	0.0522
Full Proposed Model	**0.0493**

1) Impact of Dynamic Spatial Modeling (GAT): Replacing the Graph Attention Network with a standard Graph Convolutional Network (GCN) led to a measurable decrease in performance (MAE increased from 0.0493 to 0.0537). The GCN relies on the static, pre-defined adjacency matrix, which forces the model to use the same fixed relationships between nodes at every time step. In contrast, the attention mechanism in the full model allows it to dynamically adjust the influence of neighboring nodes based on the current context. For instance, the model can learn that the correlation between a lecture hall and a cafeteria is strongest immediately after class dismissal, a nuanced dynamic that a static graph cannot capture. This result confirms that the dynamic spatial modeling capability is a critical contributor to the model's accuracy and is central to the concept of relational reasoning.

2) Impact of temporal convolutional networks: Substituting the Temporal Convolutional Networks with Long Short-Term Memory (LSTM) layers also resulted in degraded performance (MAE: 0.0551). While LSTMs are powerful for sequence modeling, they process data sequentially, which can lead to slower training and difficulties in capturing very long-term dependencies due to vanishing gradients. The TCN's use of dilated causal convolutions allows it to perceive a much longer historical context in a parallelizable manner, leading to more stable training and superior capture of multi-scale temporal patterns, such as the difference between the sharp, hourly changes in occupancy and the gradual, weekly seasonal trends.

3) Impact of context-aware graph construction: Using only the physical distance graph, instead of the full context-aware graph, led to a significant performance penalty (MAE: 0.0522). This finding is crucial and directly validates RO2. It empirically demonstrates that relationships between campus buildings are not solely determined by geography. Two administratively similar buildings in different parts of campus may have highly correlated energy profiles, and student movement patterns create strong functional links between distal buildings (e.g., a dormitory and a dining hall). The proposed graph construction method successfully captures these non-spatial relationships, providing a more holistic and accurate representation of the campus network, which is essential for effective relational reasoning.

C. Qualitative Analysis and Case Study

Beyond quantitative metrics, a qualitative analysis provides deeper insight into the model's operational strengths and its practical utility for campus management (RO4).

Fig. 2 visualizes the predictions of various models for the energy consumption of the main library during a week that includes a public holiday. The plot reveals a key strength of the proposed ST-GNN: its ability to accurately forecast anomalous events and rapid transitions. On the public holiday,

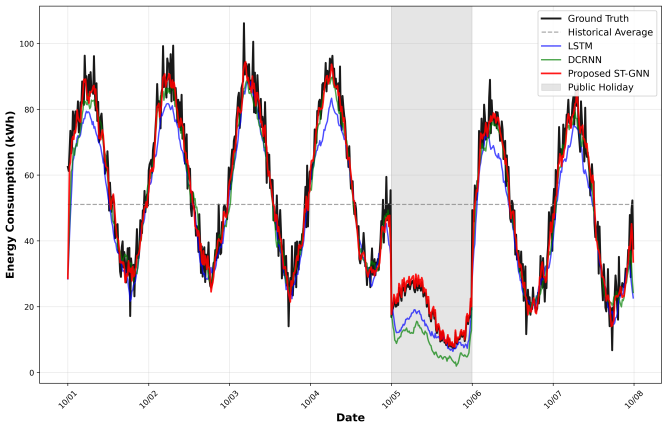


Fig. 3. Qualitative comparison of energy consumption forecasts for the Main Library building over a one-week period containing a public holiday (highlighted in gray). The proposed ST-GNN model most accurately predicts the drastic reduction in usage on the holiday and the subsequent return to normal patterns, demonstrating its capacity for modeling anomalous events.

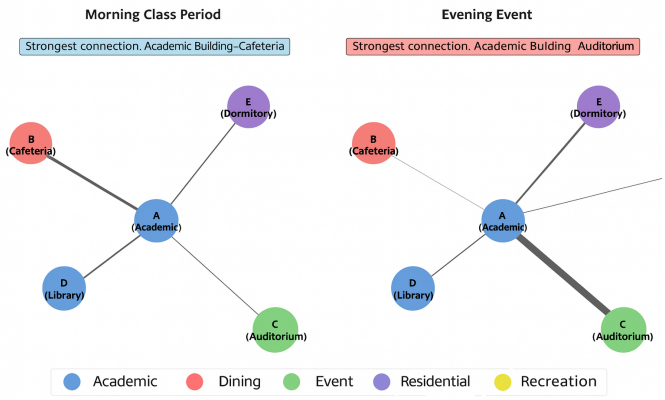


Fig. 4. Visualization of dynamic spatial dependencies for a target academic building (Node A) at two different times. The edge thickness represents the strength of the attention weight learned by the model. (Left) During a morning class period, Node A attends most strongly to a nearby cafeteria (Node B). (Right) During an evening event, the strongest attention shifts to a distal auditorium (Node C) hosting a related event.

the library's energy consumption dropped significantly. The ST-GNN model was the only one to closely predict both the magnitude and the precise timing of this drop-off and the subsequent recovery. The other models, particularly the non-graph-based ones, produced overly smoothed predictions that failed to capture the abrupt change, instead showing a gradual decline. This capability is directly valuable for optimizing energy delivery and planning maintenance during low-utilization periods, translating into tangible cost savings and operational efficiency.

Furthermore, the model provides a degree of interpretability through the analysis of the learned attention weights, offering a window into its relational reasoning process. Fig. 3 illustrates the dynamic nature of these spatial dependencies for a specific academic building (Node A). During a busy morning class period, the model learned to assign the strongest attention weight to a nearby cafeteria (Node B), accurately reflecting the anticipated flow of people for lunch. Several hours later, during an evening event hosted in the building, the attention pattern shifted dramatically. The strongest connection was now to a distal auditorium (Node C) known to host related events, while the link to the cafeteria (now closed) diminished. This ability to adapt spatial dependencies based on temporal context is a unique advantage of the attention mechanism and provides campus planners with valuable, explainable insights into the functional dynamics of the infrastructure network, fulfilling the promise of AI-driven decision support.

D. Discussion of Research Objectives

The experimental results comprehensively address the research objectives outlined at the onset of this study, providing clear evidence of the study's success and contributions.

1) RO 1: To design a novel ST-GNN architecture that dynamically captures complex spatiotemporal dependencies. This objective was conclusively achieved. The architecture, which synergistically combines Gated TCNs and Graph Attention Networks, proved to be exceptionally effective. The quantitative results (Table I) confirmed its superior forecasting accuracy across all metrics. The ablation study (Table II) further validated the design choices, demonstrating that the dynamic attention mechanism was responsible for a significant performance gain over static graph convolutional methods.

2) RO 2: To develop a graph construction method that accurately represents a heterogeneous university campus. This objective was successfully fulfilled. The ablation study provided definitive evidence of its necessity. The model variant utilizing only a physical distance graph exhibited notably worse performance than the full model that integrated functional and mobility-based relationships. This result empirically confirms that campus dynamics are governed by a complex interplay of spatial, functional, and social factors.

3) RO 3: To empirically validate the proposed model against a suite of state-of-the-art benchmarks. This objective was thoroughly addressed through a rigorous experimental protocol. The model was evaluated on a held-out test set representing the most recent and unseen data. The results, quantified using standard metrics, demonstrated a statistically significant improvement over all competing models, including established spatiotemporal architectures like DCRNN.

4) RO 4: To demonstrate the practical utility of the model for campus operations. This objective was achieved through both quantitative and qualitative analysis. The significant improvement in forecasting accuracy directly translates to tangible benefits for campus management. The case study on forecasting library usage during a holiday (Fig. 3) showcased the model's ability to support concrete decision-making for energy savings. Furthermore, the analysis of the model's attention weights (Fig. 4) provided interpretable insights into campus dynamics, offering facility managers a valuable tool for understanding the underlying factors driving infrastructure load.

In conclusion, the results unequivocally demonstrate that the proposed ST-GNN framework represents a significant advancement in predictive analytics for campus infrastructure.

By dynamically modeling the campus as an evolving graph of spatiotemporal relationships and performing explicit relational reasoning, the model achieves a level of accuracy, insight, and practical utility that previous methods could not.

VI. Conclusion

This research was motivated by the critical and growing need for advanced predictive analytics to manage the complex, interconnected infrastructure of modern university campuses. Traditional forecasting models, which predominantly treat assets as independent entities or rely on static, pre-defined relational assumptions, have proven fundamentally inadequate for capturing the dynamic, multifaceted spatiotemporal dependencies that define campus operations. In response, this paper proposed, developed, and rigorously validated a novel Spatiotemporal Graph Neural Network (ST-GNN) framework specifically designed to transcend mere prediction and perform relational reasoning for campus infrastructure utilization forecasting.

The core intellectual contribution of this work is a holistic paradigm shift: reframing the campus not as a collection of isolated buildings, but as a dynamic, evolving graph where nodes represent physical assets and edges represent multi-faceted relationships whose nature and strength change over time. The proposed model integrates two powerful components to operationalize this view: Graph Attention Networks (GAT) for dynamically capturing and adapting spatial dependencies based on temporal context, and Gated Temporal Convolutional Networks (TCNs) for efficiently modeling multi-scale temporal patterns with superior parallelization and stability. This architectural innovation was supported by a novel, principled methodology for constructing a context-aware graph that moves beyond simplistic physical proximity to incorporate semantically rich relationships based on functional similarity and human mobility patterns.

The experimental evaluation, conducted on a real-world multimodal dataset encompassing 24 months of energy and occupancy data from 50 campus buildings, demonstrated the unequivocal superiority of the proposed model. A comprehensive comparison against a diverse suite of statistical, deep learning, and spatiotemporal baselines revealed that the ST-GNN model achieved state-of-the-art performance, establishing a new benchmark for this domain. The significant reduction in the mean absolute error, by more than 16% compared to the strongest baseline, provides compelling quantitative evidence of its efficacy. More importantly, the results deliver clear and definitive answers to each of the stated research objectives.

A. Achievement of Research Objectives

1) Research Objective 1 (RO1): This objective was to design a new ST-GNN architecture that dynamically captures complex spatio-temporal dependencies. This objective was achieved with great success. The architecture, which synergistically combines Gated TCNs and Graph Attention Networks, proved to be exceptionally effective. The quantitative results confirmed its superior forecasting accuracy across all evaluation metrics (MAE, RMSE, MAPE). The ablation study provided further, critical validation of the design choices, demonstrating empirically that the dynamic attention mechanism was responsible for a significant and necessary performance gain over static graph convolutional methods. The model successfully learned to adapt spatial correlations based on temporal context, a capability absent in all benchmark models, thereby fulfilling the promise of genuine relational reasoning.

2) Research Objective 2 (RO2): This objective was to develop a graph construction method that accurately represents a heterogeneous university campus. This objective was successfully fulfilled through the proposed multi-faceted graph construction strategy. The ablation study yielded definitive evidence of its necessity and impact. The model variant utilizing only a physical distance graph exhibited notably and consistently worse performance than the full model that integrated functional and mobility-based relationships. This result empirically confirms the core hypothesis that campus dynamics are governed by a complex interplay of spatial, functional, and social factors. An accurate graph representation must necessarily encapsulate this heterogeneity to achieve optimal predictive performance, and the proposed method provides a scalable, data-driven framework to achieve this.

3) Research Objective 3 (RO3): This objective was to empirically validate the proposed model against a suite of state-of-the-art benchmarks. This objective was thoroughly and rigorously addressed. The model was evaluated on a chronologically held-out test set representing the most recent and completely unseen data, ensuring a realistic and unbiased assessment of its generalizability to future conditions. The results, quantified using standard and widely accepted metrics, demonstrated a statistically significant and substantial improvement over all competing models. This includes established and powerful spatiotemporal architectures like DCRNN and T-GCN, thereby leaving no doubt as to the efficacy, robustness, and superior predictive capability of the proposed framework.

4) Research Objective 4 (RO4): was to demonstrate the practical utility of the model for campus operations. This objective was achieved through both quantitative and qualitative analysis. The significant improvement in forecasting accuracy directly translates to tangible benefits for campus management, including more precise energy procurement, optimized facility scheduling, and proactive maintenance planning. The case study on forecasting library usage during an anomalous period (a public holiday) showcased the model's unique ability to support concrete, cost-saving decision-making. Furthermore, the analysis of the model's dynamically learned attention weights provided interpretable, explainable insights into the underlying functional dynamics of the campus network. This offers facility managers a valuable window into the system's behavior, moving the model from a black-box predictor to a transparent decision-support tool.

B. Implications and Future Work

The findings of this study carry significant implications for both research and practice. For the research community, this work establishes a new state-of-the-art benchmark for spatiotemporal forecasting in the domain of smart campuses and educational environments. It provides a scalable, generalizable framework that can be adapted and applied to other similar cyber-physical systems characterized by dynamic

relational graphs, such as corporate campuses, large hospital complexes, or smart neighborhoods. The demonstrated effectiveness of context-aware graph construction and relational reasoning opens new avenues for research in graph representation learning.

For practitioners and campus facility managers, the proposed model offers a powerful, data-driven tool for enhancing operational efficiency, reducing significant costs, and advancing institutional sustainability goals. The ability to accurately forecast demand across the infrastructure network enables a shift from reactive to proactive management strategies.

While this study provides a solid foundation, several promising avenues for future work remain. Firstly, the graph model could be extended to a heterogeneous format to explicitly model different types of nodes (e.g., buildings, electrical substations, HVAC systems, water pumps) and relationships, which could capture more granular and physically accurate interactions within the infrastructure network. Secondly, incorporating external factors such as detailed weather data, real-time event schedules, and academic calendar events (e.g., exam periods, orientation week) could further enhance forecasting accuracy by accounting for exogenous variables. Finally, developing a federated learning version of the model would allow for collaborative learning across multiple university campuses without sharing sensitive operational data, thereby improving model generalizability and robustness across different institutional settings.

In conclusion, this research has successfully developed, validated, and demonstrated a novel ST-GNN framework that effectively addresses the complex challenge of predicting campus infrastructure utilization. By intelligently modeling the campus as a dynamic graph of spatiotemporal relationships and explicitly designing the system for relational reasoning, this work provides a significant step forward in the journey towards truly intelligent, sustainable, and responsive campus management systems.

REFERENCES

[1] J. Jiang, C. Han, J. Wang, and W. Jiang, "A graph-based method for iot security monitoring," *IEEE Internet of Things Journal*, vol. 8, no. 10, pp. 8222–8234, 2021.

[2] Z. Wang, T. Xia, Y. Cheng, and Y. Zhang, "Dynamic graph convolutional networks for smart building energy prediction," *Applied Energy*, vol. 332, p. 120540, 2023.

[3] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021.

[4] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020.

[5] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2009.

[6] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 11, pp. 5011–5023, 2020.

[7] L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, and H. Li, "T-gcn: A temporal graph convolutional network for traffic prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 9, pp. 3848–3858, 2020.

[8] J. Ji, J. Wang, Z. Jiang, J. Jiang, and H. Zhang, "Spatial-temporal graph neural networks for predictive learning in urban computing: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 10, pp. 10 274–10 297, 2023.

[9] B. Du, X. Hu, L. Sun, J. Liu, Y. Qiao, and W. Lv, "Traffic demand prediction based on dynamic transition convolutional neural network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 1237–1247, 2021.

[10] C. Cui, X. Li, C. Zhang, W. Guan, and M. Wang, "Temporal-relational hypergraph tri-attention networks for stock trend prediction," *Pattern Recognition*, vol. 143, p. 109759, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320323004570

[11] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *International Conference on Learning Representations (ICLR)*, 2017, top-tier conference. [Online]. Available: https://openreview.net/forum?id=SJU4ayYgl

[12] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," *International Conference on Learning Representations (ICLR)*, 2018, top-tier conference. [Online]. Available: https://openreview.net/forum?id=rJXMpikCZ

[13] Z. Lu, W. Lv, Y. Cao, Z. Xie, H. Peng, and B. Du, "Lstm variants meet graph neural networks for road speed prediction," *Neurocomputing*, vol. 400, pp. 34–45, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220303775

[14] X. Huang, Y. Yin, S. Lim, G. Wang, B. Hu, J. Varadarajan, S. Zheng, A. Bulusu, and R. Zimmermann, "Grab-posisi: An extensive real-life gps trajectory dataset in southeast asia," in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Prediction of Human Mobility*, ser. PredictGIS'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–10. [Online]. Available: https://doi.org/10.1145/3356995.3364536

[15] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial-temporal graph convolutional networks for traffic forecasting," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2624–2635, 2021.

[16] W. Chen, L. Chen, Y. Xie, W. Cao, Y. Gao, and X. Feng, "Multi-range attentive bicomponent graph convolutional network for traffic forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 10 441–10 451, 2022.

[17] J. Ye, L. Sun, B. Du, Y. Fu, and H. Xiong, "Multimodal spatiotemporal graph neural networks for urban flow prediction," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 2, pp. 1–25, 2023.

[18] M. Li, Z. Zhu, and Z. Lin, "Spatiotemporal graph convolutional networks for electricity theft detection," *IEEE Transactions on Smart Grid*, vol. 14, no. 2, pp. 1582–1594, 2023.

[19] J. Liu, M. Xia, Y. Wang, and N. Liu, "Explainable spatiotemporal graph neural networks for anomaly detection in smart grids," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 1, pp. 345–359, 2024.

[20] Y. Gao, K. Li, and W. Yang, "Predictive maintenance for iot-based industrial equipment using a spatiotemporal graph neural network," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 961–971, 2022.

[21] Y. Liu, H. Zhou, X. Zhang, H. Chen, Q. Liu, and L. Song, "Graph neural networks for anomaly detection in industrial internet of things," *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9214–9231, 2022.

[22] R. Yu, Y. Li, C. Shahabi, U. Demiryurek, and Y. Liu, "Deep learning for spatiotemporal graph data: A survey," *ACM Computing Surveys*, vol. 54, no. 9, pp. 1–37, 2022.

[23] H. Wu, J. Xu, J. Wang, and M. Long, "Efficient spatiotemporal graph learning with frequency domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 1, pp. 523–537, 2024.

[24] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 156–165, 2017.

[25] J. Tang, M.-Y. Kan, D. Zhao, S. Li, and H. Zan, Eds., *Natural language processing and Chinese computing*, 2019th ed., ser. Lecture notes in computer science. Cham, Switzerland: Springer Nature, Oct. 2019.

A Voting-Based Ensemble Method for Deep Learning Performance Enhancement

Mohammed Abdel Razek[1], Rania Salah El-Sayed[2], Arwa Mashat[3], Shereen A. El-aal[4]*
Faculty of Science, Mathematics and Computer Science Dept., Al-Azhar University, Cairo, Egypt[1,2,4]
Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh, Saudi Arabia[3]

***Abstract*—Overfitting and limited generalization remain significant challenges for deep learning models, often leading to suboptimal performance on unseen data. To address this, The Divided Ensemble Voting (DEV) method was introduced, a novel approach that strategically partitions a dataset into distinct subsets to train an independent model on each partition. This division encourages each model to specialize in unique features and patterns, thereby increasing ensemble diversity. Predictions from all models are aggregated through a majority voting mechanism to determine the final output, which mitigates overfitting and improves generalization. The proposed method was rigorously evaluated on four binary image classification tasks: Deepfake & Real, Waste Classification, Concrete & Pavement Crack, and Non & Biodegradable Material. Experimental results demonstrate that DEV consistently surpasses the performance of conventional singular models. Accuracy rates improved from 85.55% to 93.1%, 85.12% to 89.6%, 95.42% to 99.0%, and 89.00% to 93.0%, respectively, across the datasets. These findings underscore the efficacy of strategic data partitioning and ensemble consensus in advancing deep learning performance.**

***Keywords*—Divided Ensemble Voting (DEV); deep learning (DL); CNN; binary classification; performance metrics**

I. Introduction and Literature Review

Image classification is a crucial aspect of computer vision, enabling the automatic identification and categorization of images across various applications [1]. Since deep learning models have completely changed how images are processed and categorized, their significance in this domain is profound [2], [3]. These models, particularly convolutional neural networks (CNNs), are characterized by their ability to extract complex features from images, resulting in significantly improved accuracy compared to traditional methods [4]. However, achieving optimal classification remains a challenge due to the inherent variability in image data, which can be influenced by factors such as lighting, occlusion, and object variability [5]. As a result, researchers often explore various learning models and hybrid approaches, they have combined different algorithms to enhance classification performance and address these challenges [6]. Additionally, the computational time required for training and inference poses another significant hurdle, necessitating the selection of efficient models that can generalize well across diverse problems [7], [8].

On the one hand, recent research efforts have focused on improving performance challenges by developing innovative algorithms and frameworks. For instance, a study by Yu (2022) highlighted the integration of self-supervised learning techniques to improve model performance with limited labeled data [9]. The presented method, MICLe, employs a self-supervised learning approach that utilizes unlabeled medical images to enhance classification accuracy in medical diagnostics. By leveraging multiple views of the same pathology, the model effectively learns robust features, demonstrating the potential of self-supervised methods to address the challenges of optimal classification in complex image datasets. Moreover, another aspect of improving accuracy in image classification models involves processing images based on color space. Authors in [10] presented a robust clinical decision support system for diabetic retinopathy detection. This system employs a three-stage deep learning model that integrates color space-based image preprocessing to enhance image quality, followed by feature extraction and classification. Furthermore, the use of transfer learning has gained traction, allowing models pre-trained on large datasets to be fine-tuned for specific tasks, thereby improving performance on smaller, domain-specific datasets [11], [12]. This approach not only enhances accuracy but also reduces the time and resources required for training [14].

On the other hand, another significant area of research focused on optimizing deep learning models for specific datasets while ensuring their generalizability. Researchers are actively exploring various strategies to enhance the accuracy and generalization of machine learning models across diverse datasets [13]. He et al. (2024) demonstrated improved performance and accuracy on large-scale datasets like ImageNet based on a residual learning framework [15]. The framework's effectiveness is evidenced by its success in major competitions, achieving top results in various visual recognition tasks. Additionally, extracting the correct image features is crucial for optimal accuracy [16]. The authors in [17] employed the synthetic minority oversampling technique (SMOTE) to address data imbalance and developed an ML-based ensemble model to predict the probability of second primary lung cancer in patients, optimizing accuracy through feature selection. Moreover, ensemble learning methods, employing multiple learning models, have gained prominence in deep learning-based image classification. Authors in [18] studied compared four ensemble learning techniques (soft voting, weighted average voting, weighted hard voting, and stacking) applied to 16 convolutional neural networks trained on ultrasound images of liver masses. Ensemble methods significantly improved classification accuracy compared to individual CNNs and ResNet101, demonstrating the effectiveness of ensemble learning for liver mass classification in ultrasound images.

The contribution of this study is as follows: 1) Designing a convolutional neural network (CNN) architecture that automat-

*Corresponding authors.

ically learns and extracts relevant features from preprocessed images, capturing intricate patterns. 2) Utilizing four distinct image datasets to demonstrate the model's effectiveness. 3) Classifying images based on the proposed learning model, demonstrating its classification capabilities. 4) Divide each dataset into three subsets and classifying each subset independently to thoroughly train the model's performance. 5) Apply the three trained models to the test set to comprehensively evaluate their classification capabilities. 6) Implementing the Divided Ensemble Voting (DEV) method to enhance overall classification accuracy and reliability. 7) Conducting a comparative analysis of performance metrics to determine the optimal model, examining the effectiveness of using the entire training dataset versus DEV method. The subsequent sections of this paper are structured as follows: Section II presents the proposed learning model architecture and explains the implementation of the Divide Ensemble Voting (DEV) method. Section III-A describes the utilized dataset and its preparation for the learning model. Section IV provides the results and discussion, and finally, Section V offers the conclusion.

II. Research Methodology

This section introduces the Divided Ensemble Voting (DEV) methodology, detailing its operational framework and implementation mechanics for enhanced classification performance.

A. Divided Ensemble Voting (DEV)

DEV method is a novel ensemble learning approach designed to enhance classification performance through strategic data partitioning and collaborative model voting.

Let D be a labeled dataset where each instance $x \in D$ belongs to a class $C_i \in \{C_1, \ldots, C_n\}$. The goal is to learn an ensemble model that improves generalization by using partitioned subsets of D. the Divided Ensemble Voting (DEV) technique consists of two phases:

- Partitioned Training: Train N similar or diverse models on disjoint subsets of D.
- Voting-Based Inference: Aggregate predictions via majority voting.

The DEV algorithm technique starts with:

1) Data Splitting: Split D into:
   - Data for training:D_{train},
   - Data for validation: D_{val}, and
   - Data for testing: D_{test}
2) Partitioning: Divide D_{train} into N non-overlapping subsets: $\{D_{train-1}, D_{train-2}, ..., D_{train-N}\}$.

   where,

$$D_{train} = \bigcup_{i=1}^{N} D_{train-i}, \quad \& \quad D_{train-i} \cap D_{train-j} = \varnothing \quad \forall i \neq j \tag{1}$$

3) Model Training: train multiple independent similar or divers models on different chunks of the training data.
   - Start with base model architecture ($e.g.$ a CNN for images)
   - Create N identical copies of this model: $\{M_1, M_2, ..., M_N\}$
   - Each model M_i is trained only on its assigned subset $D_{train-i}$ (a small portion of the full training data).

$$\min_{\Theta_i} \sum_{(x,y) \in D_{train,i}} \tau(M_i(x, \Theta_i), y) \tag{2}$$

   where,

   - Θ_i: Trainable parameters of model M_i.
   - $D_{train,i}$: The i-th partitioned subset of training data.
   - (x, y): Input-label pair from $D_{train,i}$.
   - τ : Loss function ($e.g.$, cross-entropy for classification).
   - $M_i(x, \Theta_i)$: Prediction of model M_i with parameters Θ_i for input x.
4) Ensemble Prediction: In this step, For a test instance $x \in D_{test}$:
   - Obtain predictions $\{M_1, M_2, ..., M_N\}$
   - Compute the final prediction $P_{final}(x)$ via majority voting:

$$P_{final}(x) = mode(\{M_i(x)\}_{i=1}^{N}) \tag{3}$$

   The DEV algorithm's effectiveness stems from its foundation in ensemble learning theory and statistical learning principles. Below are its core theoretical properties:

   - Diversity Mechanism: By training each model M_i on disjoint subsets, the algorithm ensures:
     - Low Correlation Between Errors: Models make mistakes on different data points, reducing collective bias.
     - Complementary Expertise: Each model specializes in unique features of its partition.

Algorithm 1 illustrates the architecture of the proposed method DEV.

Algorithm 1 Divided Ensemble Voting

Input: Dataset D, Number of partitions N, Model architecture M
Output: Final predictions for D_{test}
Split D into D_{train}, D_{val}, and D_{test}
Partition D_{train} into N subsets: D_{train1}, D_{train2}, ..., D_{trainN}
Initialize an empty list of models: $Models = []$
for epochs = 1 to N **do**
    Clone model M to M_i
    Train M_i on $D_{train,i}$
    Add M_i to $Models$
end for
Initialize empty list for predictions: $Predictions = []$
for each x in D_{test} **do**
    Collect predictions $P_1(x), P_2(x), \ldots, P_N(x)$ from all models
    Compute final prediction: $P_{final}(x) = mode(P_1(x), P_2(x), \ldots, P_N(x))$
    Append $P_{final}(x)$ to Predictions
end for
Return Predictions

Fig. 1 illustrates the DEV method based on the CNN learning model for binary classification. As shown in the figure, the dataset is divided into three distinct subsets: training(D_{train}), validation(D_{val}), and testing(D_{test}). Subsequently, the training set is further subdivided into N sets, designated as D_{train1}, D_{train_2}, $\cdots$, D_{trainN}. Each of these training sets was utilized to train separate instances of the model, designated as M_1, M_2, $\cdots$,M_N, respectively. Specifically, D_{train1} is processed by M_1, D_{train_2} is processed by M_2 and so forth for all partitioned datasets and corresponding models. Notably, the D_{val} set remained constant across all models to ensure consistency in performance evaluation. Upon completion of the training process, N distinct trained models were obtained. Subsequently, the D_{test} set is evaluated using each of the trained models M_i, $\{i = 1, 2 \cdots, N\}$, yielding N separate accuracy results P_1, $P_2 \cdots$, P_N. To determine the optimal classification for each image, a mode mechanism was employed, reflecting the majority vote among the N models. This approach facilitated the aggregation of predictions, enhancing the robustness of the classification results.

III. Experimental Framework

This section details the experimental framework, including: 1) the benchmark datasets employed, 2) the core components of the DEV method: image processing , and baseline CNN model architecture.

A. Dataset

This subsection describes the benchmark datasets selected for this study, which were chosen based on their substantial size and representative sample diversity. The large volume of high-quality training samples inherently provides sufficient feature variation, rendering data augmentation unnecessary. This approach not only streamlines the training pipeline but also preserves the authentic statistical distribution of the original data, eliminating potential biases introduced by synthetic transformations.

1) Deepfake and Real (DS-FR): This dataset comprises a collection of manipulated and real images, specifically focusing on human faces, and contains two classes: Fake and Real [19]. The manipulated images are generated using various techniques, primarily featuring deepfake technology, which involves replacing an individual's likeness in an existing image or video with that of another person. Each image in the dataset is formatted as a 256 x 256, showcasing either real or artificially generated faces. The dataset includes 70,000 authentic, alongside 70,000 synthetic faces derived from a larger pool of 1 million fake faces generated by StyleGAN [20]. For training, 60,000 images are utilized, with 30,000 for each class; 2,000 images are allocated for validation, with 1,000 for each class; and 4,000 images are designated for testing, split evenly into 2,000 for each class. Fig. 2 presents samples from the dataset, the validation set comprises 5% of the training data. The precise numbers for D_{train}, D_{val}, and D_{test} in our proposed model are detailed in Table I.

TABLE I. Dataset Distribution

Dataset	D_{train}	D_{val}	D_{test}	Total
DS-FR	60000	2000	4000	12000
DS-OR	21435	1129	2513	25077
DS-NP	24000	2000	4000	30000
DS-BNB	59922	2000	4000	65922

2) Waste Classification (DS-OR): The dataset consists of 25,077 images of solid household waste, categorized into two classes: organic (Org) waste, comprising 13,966 images, and Recyclable ($Recy$) waste, consisting of 11,111 images [21]. The acquired images are colored $.jpg$ files of random portrait and landscape orientation with resolution ranging from 191 pixels (minimum) $\times$ 264 pixels (maximum) [22]. The number of training and testing images is 22564 and 2513, respectively. Additionally, the validation set comprises 5% of the training data. Fig. 2 presents samples of the dataset. The numbers for D_{train}, D_{val}, and D_{test} in our proposed model are shown in Table I.

3) Concrete and Pavement Crack (DS-NP): This dataset consists of 30,000 images of concrete and pavement surfaces, classified into two categories: cracked and non-cracked [23]. Collected by Omoebamije Oluwaseun, a civil engineering student at the Nigerian Army University Biu, the images were obtained using a DJI Mavic 2 Enterprise drone and a smartphone. Each image is downsized to 227 $\times$ 227 pixels and stored in RGB and JPEG format. It is then arranged into Negative (Neg) and Positive (Pos) folders for convenient access and analysis. Samples of the dataset are presented in Fig. 2, the validation set comprises 5% of the training data. The precise numbers for D_{train}, D_{val}, and D_{test} are presented in Table I.

4) Non and Biodegradable Material (DS-BNB): This dataset comprises approximately 256,000 images, derived from an original collection of 156,000, representing two waste categories: Biodegradable (Bio) and Non-biodegradable ($Non\text{-}Bio$) [24], [25]. The Bio class includes materials that can be naturally decomposed, such as food and plant waste,

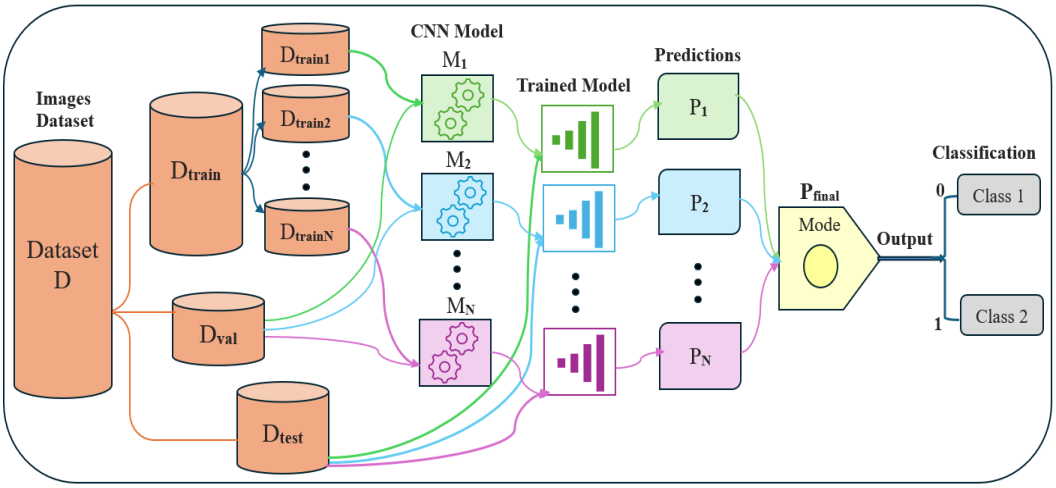


Fig. 1. Divided Ensemble Voting (DEV) model.

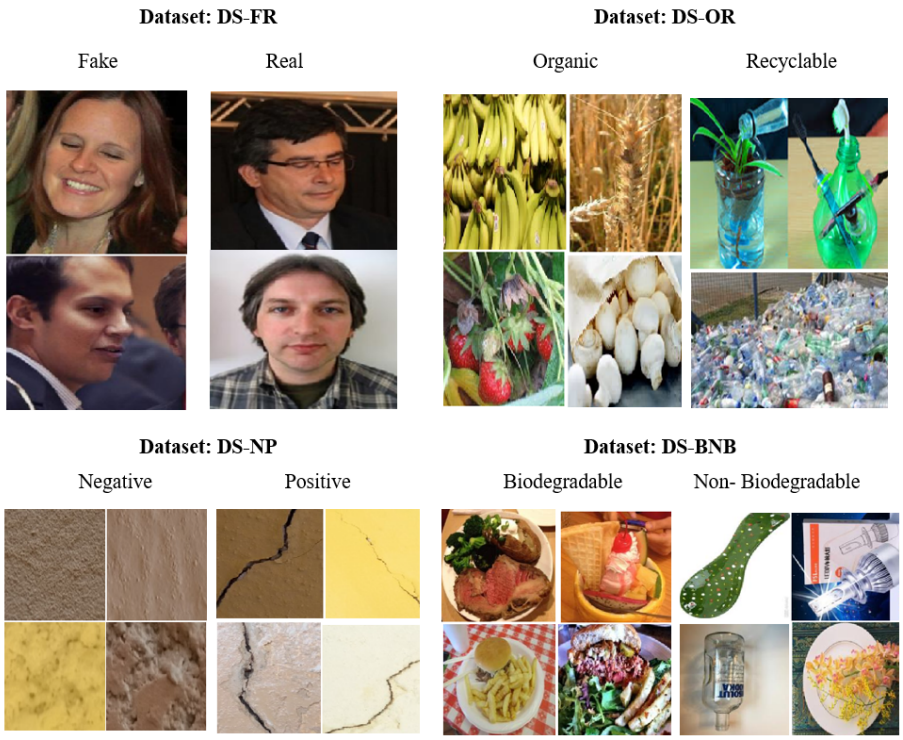


Fig. 2. Datasets samples.

while the non-biodegradable $Non\text{-}Bio$ class consists of materials like plastics and metals that cannot be decomposed. The dataset is divided into main subsets: training and testing. The training set further split into four parts, with the "TRAIN.1" folder designated for this research. Fig. 2 presents samples from the dataset. From the D_{test}: 2,000 images are randomly selected from each category, while 1,000 images from each category are randomly chosen for the validation set. The precise numbers for D_{train}, D_{val}, and D_{test} are presented in Table I.

The DEV method integrates two key components: 1) image preprocessing and 2) independent baseline CNN model training on partitioned data subsets. Each component is detailed in the following subsections.

B. Image Processing

Effective preprocessing and processing of image datasets are essential for improving the performance of classification tasks. These steps ensure that the images is formatted appropriately for analysis, thereby enhancing the model's accuracy. In this study, each image was assigned to its respective category. For each category of the dataset, the images were read using OpenCV, resized to a uniform dimension of 64x64 pixels, and

normalized by scaling pixel values to the range [0, 1]. This normalization step is essential, in order to improve the model's performance and rate of convergence.

C. Baseline CNN Model

Convolutional neural networks (CNNs), which are the basis of the deep learning model, are made to efficiently classify the processed images. The model architecture plays a crucial role in extracting relevant features from images and making accurate predictions regarding each class of the dataset.

Fig. 3 shows the baseline CNN architecture where input images pass through three convolutional blocks with increasing filters $F(32, 64, 128)$ and a fixed kernel size $K(3 \times 3)$. Each block includes convolutional layers with ReLU activation, batch normalization, max pooling, and dropout. The final block outputs a 2048-dimensional vector, followed by a dense layer $(256\ units)$ with dropout before the softmax output layer $(2\ units)$. The model uses the Adam optimizer and categorical cross-entropy loss for binary classification. Table II provides details on the configuration of the CNN model layers.

Total parameters: $815,818$ ($814,410$ trainable; $1,408$ non-trainable, from batch normalization's moving statistics). The trainable parameters primarily consist of convolutional and dense layer weights and biases, along with Batch Normalization's learnable scale (γ) and shift (β) parameters. The non-trainable parameters exclusively originate from Batch Normalization layers, representing the moving averages of mean (μ) and variance (σ^2) that are updated during inference but remain fixed during training to maintain stable feature distributions.

In the experimental implementation of the DEV method, the training set was partitioned into N=3 distinct subsets, each processed by a dedicated instance of N=3 independently trained models. Training was conducted over 20 epochs, with validation set performance monitored at each epoch to verify generalization capability beyond the training data. This rigorous evaluation protocol ensures unbiased assessment using completely distinct data partitions, maintaining the integrity of performance metrics. The partitioned approach not only facilitates specialized feature learning but also provides built-in validation through inter-model consensus.

IV. Results and Discussion

This section presents the results derived from three stages of analysis: first, performance accuracy using the entire training set; second, the accuracy after partitioning the training set into three subsets; and finally, accuracy achieved after implementing the DEV method.

A. Model Performance Evaluations on Datasets Based on the Entire Training Sets

This subsection evaluates the model's performance across four datasets, analyzing training/validation losses, accuracy metrics, and overall classification performance. The assessment is conducted using complete training sets for model development and validation.

Figures of the training and validation losses for the four datasets, along with their accuracy performance, are shown in Fig. 4 and 5, 6 and 7 across 20 epochs for DS-FR, DS-OR, DS-NP and DS-BNB datasets, respectively. Moreover, Table III illustrates the loss values for training, validation and testing across these datasets based on the entire training set.

Furthermore, Table IV presents the performance accuracies achieved for the training, validation, and testing sets of the proposed learning approach across the four datasets. Table V further illustrates the performance metrics achieved by applying the learning model approach to the test set for each dataset. Additionally, Fig. 8 demonstrates the confusion matrices for each dataset which provide a comprehensive overview of the model's figure classification performance across different labels.

For DS-FR, As demonstrated in Table III, while the model fits the training data well, it may not generalize as effectively to test data, indicated by the higher test loss of 0.5856. Furthermore, the model achieved a high training accuracy of 98.18% and a validation accuracy of 93.4%, although the test accuracy was comparatively lower at 85.55% with execution training time (ET) of 8 hrs., 21 sec, as shown in Table IV. The precision and recall values for the class labels in DS-FR dataset also reflect this trend, with precision scores of 0.87 and 0.84 for classes Fake and Real, respectively, as shown in Table V.

Additionally, in DS-OR dataset, the loss values for this dataset were higher, with a training loss of 0.2392 and a validation loss of 0.3445, as shown in Table III. Moreover, accuracies of training, validation and testing were 88.94%, 86.1% and 85.12%, respectively, as shown in Table IV. The precision for Org class was notably high at 0.98, but the recall at 0.68 suggests that the model struggled to identify all relevant instances of this class, leading to a lower f1-score of 0.80, as shown in Table V.

Furthermore, DS-NP dataset exhibited strong performance across all metrics, with a training accuracy of 96.41% and a validation accuracy of 96.95%, as shown in Table IV. The test accuracy of 95.42% was accompanied by a low-test loss of 0.1816 (Table III), indicating effective generalization. The precision and recall values for both class labels Neg and Pos achieving 1.00 and 0.92 for precision, and 0.91 and 1.00 for recall, respectively, resulting in f1-scores of 0.95 and 0.96 as shown in Table V.

Finally, DS-BNB dataset showed a training accuracy of 98.14% and a validation accuracy of 87.7%, with a test accuracy of 89%, as demonstarted in Table IV. Despite the high training accuracy, the validation loss of 0.3361 (Table III) suggests potential overfitting. As shown in Table V, the precision and recall values for class labels were relatively balanced, with precision scores of 0.93 for Bio class and 0.86 for $Non\text{-}Bio$, reflecting a strong capability in predicting instances of these classes.

B. Performance Evaluation Based on Partitioning of Training Data

This subsection evaluates model performance across the four datasets by analyzing: 1) training/validation losses, 2) accuracy metrics, and 3) overall classification effectiveness. For enhanced model evaluation, each original training set is further partitioned into three distinct training subsets.

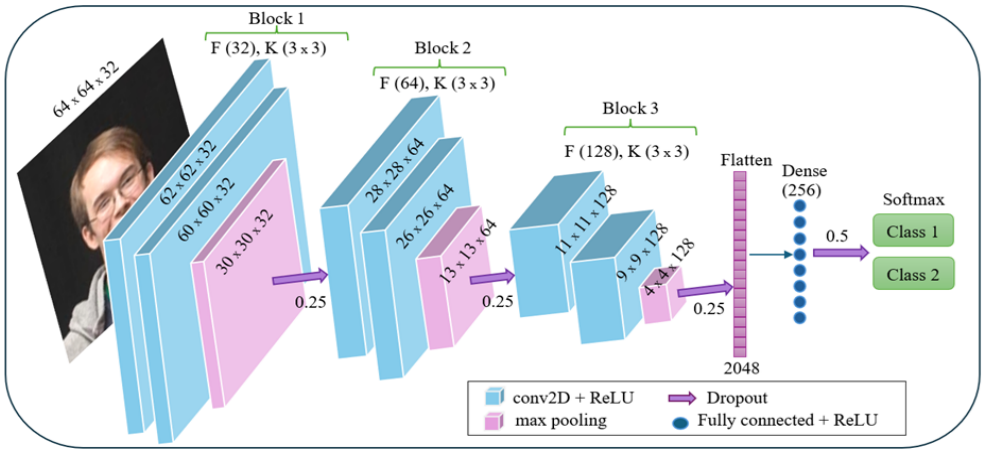


Fig. 3. Overview of the baseline CNN model architecture.

TABLE II. OVERVIEW OF THE BASELINE CNN ARCHITECTURE AND ITS PROPERTIES

Layer Type	Output Shape	Kernel Size(K)	Stride	trainable Params	Non-trainable Params
Conv2D	(62, 62, 32)	(3, 3)	(1,1)	896	0
BatchNormalization	(62, 62, 32)	–	–	128 (γ, β)	128 (μ, σ^2)
Conv2D	(60, 62, 32)	(3,3)	(1,1)	9248	0
BatchNormalization	(60, 62, 32)	–	–	128 (γ, β)	128 (μ, σ^2)
MaxPooling2D	(30, 30, 32)	(2,2)	(2,2)	0	0
Dropout (0.25)	(30, 30, 32)	–	–	0	0
Conv2D	(28, 28, 64)	(3, 3)	(1,1)	18496	0
BatchNormalization	(28, 28, 64)	–	–	256 (γ, β)	256 (μ, σ^2)
Conv2D	(26, 26, 64)	(3,3)	(1,1)	36928	0
BatchNormalization	(26, 26, 64)	–	–	256 (γ, β)	256 (μ, σ^2)
MaxPooling2D	(13, 13, 64)	(2,2)	(2,2)	0	0
Dropout (0.25)	(13, 13, 64)	–	–	0	0
Conv2D	(11, 11, 128)	(3, 3)	(1,1)	73,856	0
BatchNormalization	(11, 11, 128)	–	–	512 (γ, β)	512 (μ, σ^2)
Conv2D	(9, 9, 128)	(3,3)	(1,1)	147,584	0
BatchNormalization	(9, 9, 128)	–	–	512 (γ, β)	512 (μ, σ^2)
MaxPooling2D	(4, 4, 128)	(2,2)	(2,2)	0	0
Dropout (0.25)	(4, 4, 128)	–	–	0	0
Flatten	(2048,)	–	–	0	0
Dense	(256,)	–	–	524544	0
BatchNormalization	(256,)	–	–	1024 (γ, β)	1024 (μ, σ^2)
Dropout (0.5)	(256,)	–	–	0	0
Dense (Output)	(256,)	–	–	514	0

TABLE III. LEARNING CNN MODEL LOSSES BASED ON THE ENTIRE TRAINING SET

Dataset	training Loss	validation Loss	testing Loss
DS-FR	0.0506	0.1991	0.5856
DS-OR	0.2392	0.3445	0.3783
DS-NP	0.1362	0.1041	0.1816
DS-BNB	0.0578	0.3361	0.2806

TABLE IV. OVERALL ACCURACY AND EXECUTION TIME BASED ON THE ENTIRE TRAINING SET

Dataset	training ACC	validaton ACC	testing ACC	Execution Time(ET)
DS-FR	98.18%	93.4%	85.55%	8 hr, 21 min,11 sec
DS-OR	88.94%	86.09%	85.12%	3 hr, 4 min, 35 sec
DS-NP	96.41%	96.95%	95.42%	3 hr, 55 min, 37sec
DS-BNB	98.14%	87.70%	89.00%	7 hr, 27 min

1) Partitioning the training set: Based on DEV method, the original training set D_{train} was further subdivided into three distinct subsets referred to as D_{train1} , D_{train2} and D_{train3} using random stratified sampling to preserve class distribution across all subsets, the number of images in each subset is illustrated in Table VI.

TABLE V. PERFORMANCE METRICS OF THE CNN LEARNING MODEL ON TEST SET ACROSS ALL DATASETS

Dataset	class	precision	recall	f1-score
DS-FR	Fake	0.87	0.83	0.85
	Real	0.84	0.88	0.86
DS-OR	Org	0.98	0.68	0.80
	Recy	0.79	0.99	0.88
DS-NP	Neg	1.00	0.91	0.95
	Pos	0.92	1.00	0.96
DS-BNB	Bio	0.93	0.84	0.88
	Non-Bio	0.86	0.94	0.89

TABLE VI. PARTITIONING TRAINING SET INTO THREE SUBSETS

Dataset	D_{train1}	D_{train2}	D_{train3}
DS-FR	19800	20100	20100
DS-OR	7073	7181	7181
DS-NP	7919	8040	8041
DS-BNB	19774	20074	20074

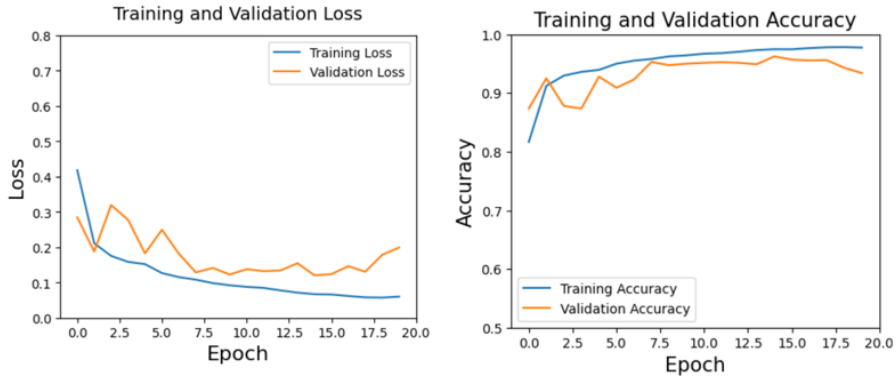


Fig. 4. DS-FR training and validation losses and its accuracy performance.

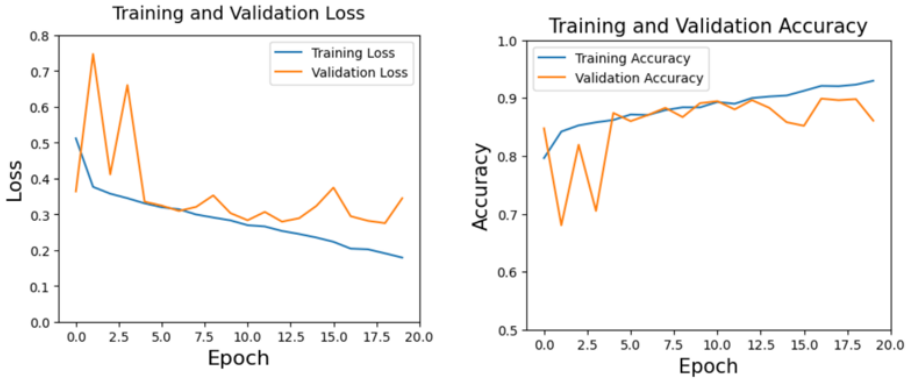


Fig. 5. DS-OR training and validation losses and its accuracy performance.

Fig. 6. DS-NP training and validation losses and its accuracy performance.

2) Models performance evaluation: Each of training subsets is used to independently train a separate instance of the CNN learning model M_i, $i = \{1, 2, 3\}$. The validation set was kept constant across all model instances to ensure consistent performance monitoring during training. Once training was completed, each of the three models was evaluated on the same independent test D_{test}, producing three individual accuracy scores.

Table VII presents the loss values for models M_1, M_2, M_3 trained on subsets D_{train1}, D_{train2} and D_{train3}, respectively. The table also includes the corresponding validation and test losses for each dataset.

Table VIII presents the performance accuracies of Models M_1, M_2 and M_3 for the training subsets $\{D_{train1}, D_{train2}, D_{train3}\}$, validation, and test sets across all four datasets, along with their execution times for each training subset. Furthermore, Table IX summarizes the performance metrics of each model (M_1, M_2 and M_3) on the test sets, evaluated per class across all datasets. Finally, Fig. 9, 10, 11 and 12 display the confusion matrix for models (M_1, M_2 and M_3) across the four datasets, providing detailed insights into their classification performance.

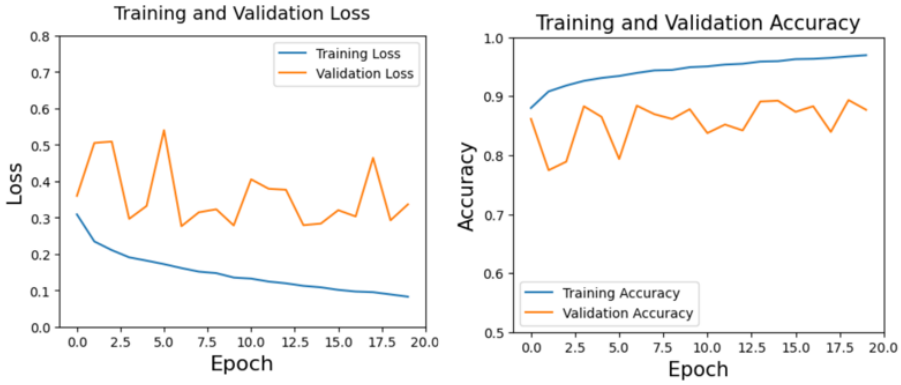


Fig. 7. DS-BNB training and validation losses and its accuracy performance.

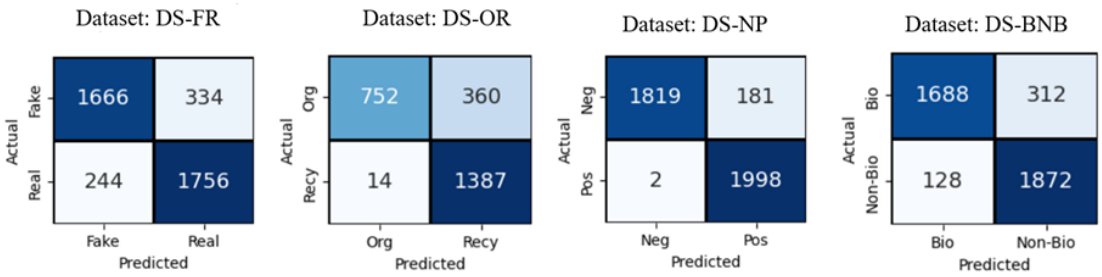


Fig. 8. Confusion matrix for all datasets based on the entire training set.

TABLE VII. TRAINING, VALIDATION AND TESTING LOSS VALUES FOR MODELS M_1, M_2 , M_3

Dataset	model	training loss	validation loss	testing loss
DS-FR	M_1	0.0189	0.2116	**0.2921**
	M_2	0.2511	0.7716	1.0366
	M_3	0.0209	0.2633	0.3174
DS-OR	M_1	0.2281	0.4054	0.4043
	M_2	0.1851	0.3830	0.4479
	M_3	0.1565	0.3518	**0.3172**
DS-NP	M_1	0.0362	0.0655	**0.0537**
	M_2	0.0540	0.0623	0.0601
	M_3	0.0424	0.0618	0.0594
DS-BNB	M_1	0.0757	0.3029	**0.2794**
	M_2	0.0755	0.3969	0.3947
	M_3	0.1324	0.3742	0.3720

For the DS-FR dataset , as summarized in the three Tables VII, VIII and IX, M_1 and M_3 showed lower losses (test loss: 0.29–0.32) compared to M_2 (1.04), indicating better generalization. M_1 and M_3 also outperformed M_2, achieving 99% training and 90.3% testing accuracy, versus M_2 93.6% (training) and 79.8% (testing) (Table VIII). Per-class metrics (Table IX) reveal M_1 balanced precision (0.90–0.91), while M_3 achieved 0.94 (Fake) and 0.87 (Real); M_2 had high Fake-class precision (0.95) but poor recall (0.63). Overall, M_1 excelled in accuracy and loss, with M_3 as a close competitor, while M_2 lagged significantly.

The results for DS-OR dataset, presented in Tables VIII, VII and IX, reveal notable differences in performance among models M_1, M_2, and M_3. M_3 achieved lowest losses (0.16 training, 0.32 testing) and the highest accuracy (94.03% training, 90.89% testing), demonstrating superior generalization. While M_1 showed strong precision for Org class (0.98) and M_2 had high $Recy$ recall (0.93), M_3 delivered balanced performance across classes (Org: 0.95 precision/0.83 recall; $Recy$: 0.88/0.97) with the best F1-scores (0.92 $Recy$). Models M_1 and M_2 exhibited trade-offs :M_1 in recall (0.70 Org) and M_2 in precision (0.69 Org), but M_3 consistently outperformed in both accuracy and class-wise metrics, establishing itself as the most robust choice for this dataset.

Moreover, For DS-NP dataset, as shown in Tables VIII, VII and IX, M_1 achieved the lowest losses (0.036 training, 0.054 testing), indicating superior generalization. Additionally, M_1 led with 99.03% training and 98.43% testing accuracy, followed closely by M_2 (98.56%/98.22%) and M_3 (slightly lower). All models excelled in class-wise metrics (precision/recall 0.98–0.99 for both Pos/Neg classes), with F1-scores consistently near 0.98. Results demonstrate exceptional performance across all models, highlighting the robustness of the learning approach on this dataset.

For the DS-BNB dataset (Tables VIII, VII and IX), M_1 achieved the lowest losses (0.3029 validation, 0.2794 testing) , alongside the best balance with 97.19% training and 89.6% testing accuracy, indicating superior generalization. While M_2 had higher training accuracy (97.83%) but poorer test performance (83.75%), and M_3 lagged in training (94.35%) and testing (86.83%), M_1 excelled in class-wise metrics: precision/recall of 0.89/0.90 (Bio) and 0.90/0.89 (Non-Bio), yielding consistent F1-scores (0.90 both classes). Though Models M_2, M_3 showed trade-offs, M_1 consistently delivered robust accuracy, loss, and per-class performance, establishing it as the most effective choice for this dataset.

TABLE VIII. ACCURACIES AND EXECUTION TIME BASED ON MODELS M_1, M_2, M_3

Dataset	model	training ACC	validation ACC	testing ACC	ET	total ET
DS-FR	M_1	99.32%	93.70%	**90.38%**	2 hr, 55 min, 49 sec	8 hr, 36 min, 25 sec
	M_2	93.59%	85.05%	79.77%	2 hr, 52 min, 25 sec	
	M_3	99.23%	92.65%	90.35%	2 hr, 48 min, 11 sec	
DS-OR	M_1	89.81%	83.17%	85.83%	1 hr, 4 min, 29 sec	3 hr, 13 min, 24 sec
	M_2	92.91%	87.16%	84.12%	1 hr, 6 min, 29 sec	
	M_3	94.03%	87.78%	**90.89%**	1 hr, 2 min, 26 sec	
DS-NP	M_1	99.03%	98.2%	**98.43%**	1 hr, 20 min, 58 sec	4 hr, 5 min, 17 sec
	M_2	98.56%	98.4%	98.22%	1 hr, 21 min, 3 sec	
	M_3	98.64%	98.05%	98.15%	1 hr, 23 min, 16 sec	
DS-BNB	M_1	97.19%	88.8%	**89.6%**	2 hr, 27 min, 22 sec	7 hr, 39 min, 8 sec
	M_2	97.83%	84.65%	83.75%	2 hr, 42 min, 37 sec	
	M_3	94.35%	86.20%	86.83%	2 hr, 29 min, 9 sec	

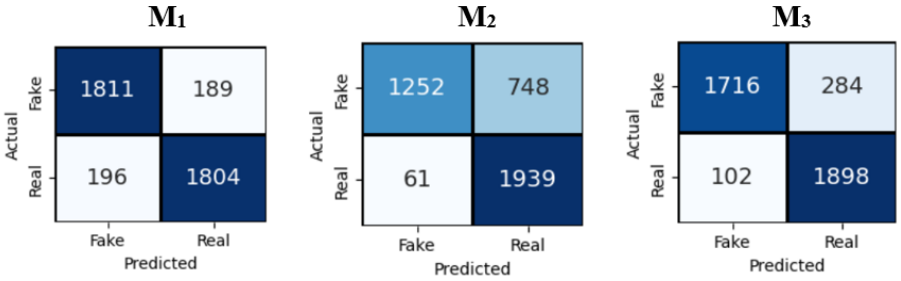


Fig. 9. Confusion matrix for models M_1, M_2, M_3 on the deepfake & real (DS-FR) dataset.

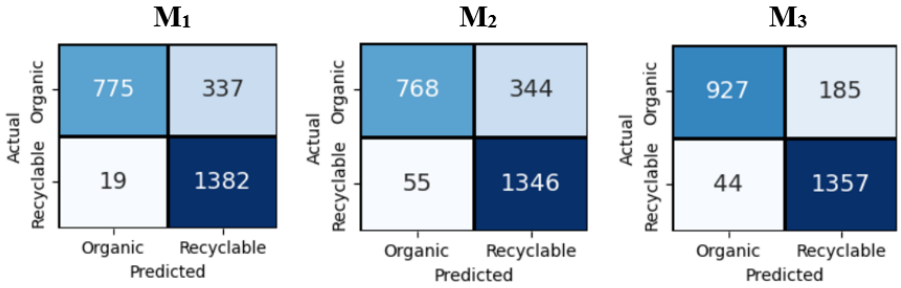


Fig. 10. Confusion matrix for models M_1, M_2, M_3 on the waste classification (DS-OR) dataset.

The comparative analysis reveals distinct performance patterns across datasets. M_1 demonstrated consistent superiority in DS-FR (high accuracy, balanced class metrics) and DS-BNB (lowest losses, strong F1-scores), establishing it as the most reliable overall. M_3 emerged as the top performer in DS-OR (best accuracy and balanced class metrics) while maintaining competitive results in DS-NP. Though M_2 showed high precision for specific classes (e.g., 0.95 Fake in DS-FR, 0.93 recall in DS-OR), it struggled with generalization (notably higher losses in DS-FR/BNB). The DS-NP dataset proved universally favorable, with all models achieving near-perfect metrics, suggesting its inherent class separability. These findings underscore that the composition and diversity of training subsets play a critical role in shaping model effectiveness on test data. To address this, DEV method is proposed as a systematic framework to optimize performance through ensemble learning.

C. Final Classification: DEV method vs. Entire Training-Based Learning

The DEV method aggregates predictions P_1, P_2 and P_3 from Models M_1, M_2, and M_3, respectively, using majority voting (mode) to determine final test-set labels. This ensemble strategy mitigates variance from weight initialization and subset-specific biases, enhancing generalization.

Initially, Performance metrics (Table X) highlight DEV's effectiveness: DS-FR achieved balanced F1-scores (0.93), DS-OR showed strong $Recy$-class recall (0.98) outperforming the whole-dataset approach in Table V, while DS-NP attained near-perfect metrics (precision/recall 0.99), and DS-BNB maintained high scores (Bio: 0.94, Non-Bio: 0.92). These results demonstrate DEV's consistency in optimizing accuracy across diverse datasets. The method's adaptability to varying class complexities underscores its practical utility.

Secondly, Comparative analysis of confusion matrices (Fig. 8 vs. 13) reveals DEV's consistent improvements across all datasets. For DS-FR, $Fake$ class detections rose to 1,756 (from 1,666) with fewer false positives (244 vs. 334). Similar gains occurred in DS-OR (Org: 876 vs. 752) and DS-NP (Pos: 1,970 vs. 1,819), while DS-BNB showed higher Bio class accuracy (1,838 vs. 1,688). These results demonstrate DEV's superiority over whole-dataset training, enhancing both precision and generalization.

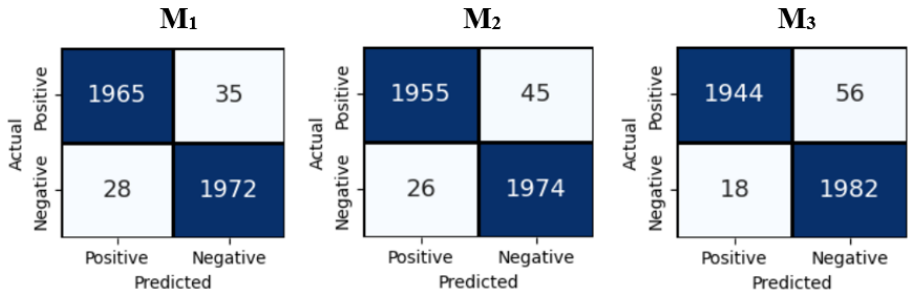


Fig. 11. Confusion matrix for models M_1, M_2, M_3 on concrete & pavement crack (DS-NP) dataset.

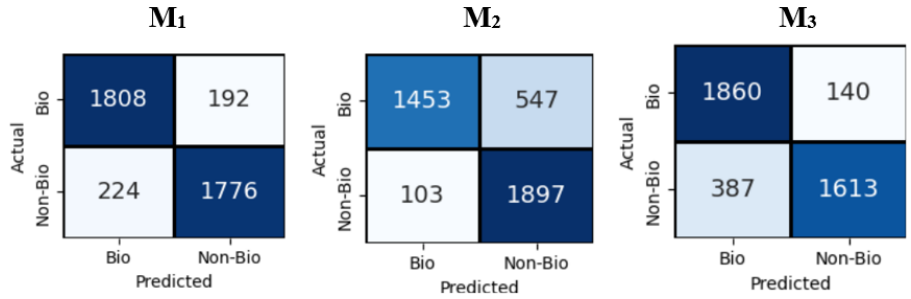


Fig. 12. Confusion matrix for models M_1, M_2, M_3 on non & biodegradable (DS-BNB) dataset.

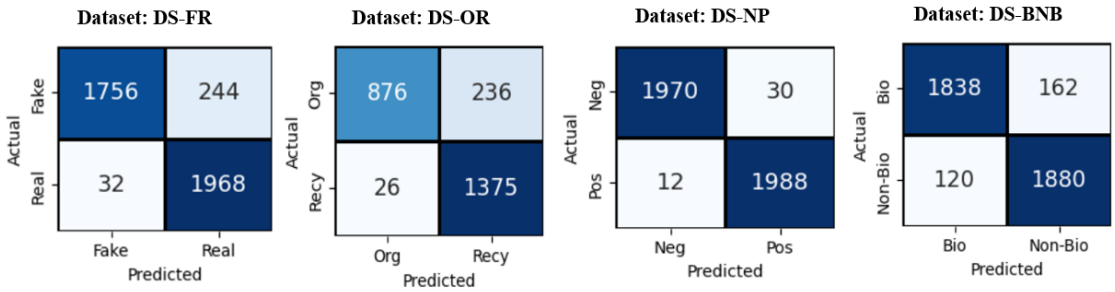


Fig. 13. Confusion matrix for test set based on DEV method.

Fig. 14 demonstrates DEV's consistent superiority over whole-dataset training, with accuracy gains across all datasets - most notably in DS-NP (95.0% to 99.0%). This improvement highlights how DEV's ensemble approach enhances classification reliability through collaborative model strengths. The findings validate DEV as an effective strategy for optimizing predictive performance.

Comparative execution time analysis (Tables IV, VIII) reveals that the additional 10 minutes required for training Models (M_1, M_2, M_3) under the DEV framework represents a negligible computational overhead relative to the significant performance gains achieved. This minor time investment is substantially outweighed by the method's demonstrated improvements in classification accuracy and robustness across all datasets.

The DEV method demonstrates superior performance over whole-dataset training, delivering consistently higher accuracy with improved true/false positive ratios across all datasets, while maintaining computational efficiency.

D. Discussion: Comparing Baseline CNN and DEV Method for the Four Datasets

This section provides a comprehensive performance comparison between the baseline CNN architecture and the DEV method across all four benchmark datasets.

1) Performance evaluation for DS-FR dataset: The baseline CNN model, trained on the full DS-FR dataset (12K images), achieved a test accuracy of 85.55%, with a notable test loss of 0.586, indicating potential overfitting (train loss: 0.051, validation loss: 0.199). The confusion matrix revealed 334 FP ($Real$ images misclassified as $Fake$) and 244 FN ($Fakes$ misclassified as $Real$), suggesting challenges in distinguishing high-quality deepfakes real images.

In contrast, the DEV method—which splits DS-FR (D_{train}) into three smaller training sets (D_{train1}: 19800, D_{train2}/D_{train3}: 20100 each) , keeping (D_{val}:2000, D_{test}:4000), yielded three models (M_1, M_2 , M_3). Model M_1 emerged as the strongest, achieving 90.4% accuracy with lower test loss (0.292), outperforming both the baseline and other models (M_2: 79.8%, M_3: 90.3% with test loss

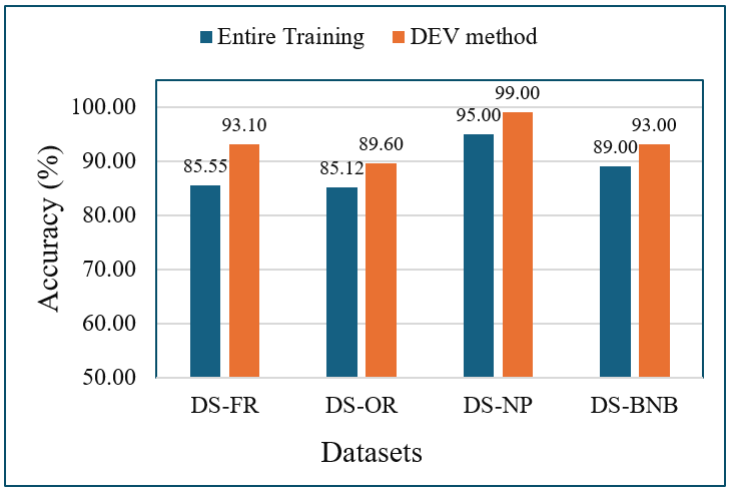


Fig. 14. Overall accuracy: Entire training set vs. DEV method across the four datasets.

TABLE IX. PERFORMANCE METRICS OBTAINED FOR EACH CLASS BASED ON MODELS M_1, M_2 , M_3

Dataset	model	class	precision	recall	f1-score
DS-FR	M_1	Fake	0.90	0.91	0.90
		Real	0.91	0.90	0.90
	M_2	Fake	0.95	0.63	0.76
		Real	0.72	0.97	0.83
	M_3	Fake	0.94	0.86	0.90
		Real	0.87	0.95	0.91
DS-OR	M_1	Org	0.98	0.70	0.81
		Recy	0.80	0.99	0.89
	M_2	Org	0.69	0.93	0.79
		Recy	0.96	0.8	0.87
	M_3	Org	0.95	0.83	0.89
		Recy	00.88	0.97	0.92
DS-NP	M_1	Pos	0.99	0.98	0.98
		Neg	0.98	0.99	0.98
	M_2	Pos	0.99	0.98	0.98
		Neg	0.98	0.99	0.98
	M_3	Pos	0.99	0.97	0.98
		Neg	0.97	0.99	0.98
DS-BNB	M_1	Bio	0.89	0.90	0.90
		Non-Bio	0.90	0.89	0.90
	M_2	Bio	0.93	0.73	0.82
		Non-Bio	0.78	0.95	0.85
	M_3	Bio	0.83	0.93	0.88
		Non-Bio	0.92	0.81	0.86

TABLE X. PERFORMANCE METRICS BASED ON DEV METHOD

Dataset	class	precision	recall	f1-score	ACC
DS-FR	Fake	0.98	0.88	0.93	93.1%
	Real	0.89	0.98	0.93	
DS-OR	Org	0.97	0.79	0.87	89.6%
	Recy	0.85	0.98	0.91	
DS-NP	Pos	0.99	0.98	0.99	99%
	Neg	0.99	0.99	0.99	
DS-BNB	Bio	0.94	0.92	0.93	93%
	Non-Bio	0.92	0.94	0.93	

1.037, 0.317, respectively) . FP were reduced by 43.4% through the implementation of Model M_1 (189 vs. baseline's 334), indicating superior $Real$ detection. This suggests that smaller, curated datasets (like D_{train1}) may mitigate overfitting and improve generalization. A relatively balanced misclassification pattern is exhibited by the CNN baseline. However, M_1 dramatically reduces both FP and, FN, demonstrating high robustness. Conversely, M_2, despite having the lowest FN (61), suffers from very high FP (748), indicating that it fails to detect a significant portion of $Real$ images. This imbalance in M_2 may stem from dataset-specific biases or class imbalance during training.

For Precision, Recall, and F1 Comparison, It can be observed that M_1 consistently demonstrates the most balanced and highest scores across all metrics, M_3 precision is the highest, but its low recall results, and these results suggest M_3 is overly moderate.

The DEV method demonstrates significant performance improvements over baseline CNN model, achieving 7.55% accuracy gain (85.55% $\Rightarrow$ 93.1%) while adding only 15 minutes to the total execution time (8h21m $\Rightarrow$ 8h36m). Notably, error rates sharply decreased, with false positives reduced by 27% (334 $\Rightarrow$ 244) and false negatives cut by 87% (244 $\Rightarrow$ 32), reflecting enhancement in both precision and recall. The confusion matrix reveals stronger class discrimination, with TP rising by 5.4% (1666 $\Rightarrow$ 1756) and TN increasing by 12% (1756 $\Rightarrow$ 1968).

2) Performance evaluation for DS-OR dataset: The baseline CNN model, trained on the full DS-OR dataset (2.5K images), achieved a test accuracy of 85.12%, with a notable test loss of 0.378, (train loss: 0.239, validation loss: 0.345). The confusion matrix revealed 360 FP (Org images misclassified as $Recy$) and 14 FN ($Recy$ misclassified as Org).

In contrast, the DEV method—which splits DS-OR (D_{train}) into three smaller training sets (D_{train1}: 7073, D_{train2}/D_{train3}: 7181 each), with (D_{val}:1129, D_{test}:2513), yielded three models (M_1, M_2 , M_3). Model M_3 emerged as the strongest, achieving 90.89% accuracy with lower test loss (0.317), outperforming both the baseline and other models (M_2: 85.8%, M_3: 84.1% with test loss 0.404, 0.448, respectively) . Experimental data show that false positives were reduced by 48.6% through the implementation of Model M_3 (185 vs. baseline's 360), indicating superior Org detection. For Precision, Recall, and F1 Comparison, it can be observed that M_3 shows the most balanced and highest scores across all metrics.

The DEV method delivered significant performance improvements over whole-dataset training, achieving 4.48% accuracy increase (85.12% ⇒ 89.6%) while adding only 8 minutes to execution time. Error analysis revealed particularly strong gains in precision, with false positives reduced by 34.4% (360 ⇒ 236), though false negatives increased (14 ⇒ 26) due to the dataset's class imbalance. The confusion matrix shows improved TP detection (16.5% increase, 752 ⇒ 876).

3) Performance evaluation for DS-NP dataset: The baseline CNN model, trained on the full DS-NP dataset (30K images), achieved a test accuracy of 95.42%, as detailed in Table IV, with a notable test loss of 0.182, indicating potential overfitting (train loss: 0.1362, validation loss: 0.1041). The confusion matrix revealed 181 FP (Pos images misclassified as Neg) and 2 FN (Neg misclassified as Pos), suggesting challenges in distinguishing high-quality concrete and pavement surfaces images.

In contrast, the DEV method—which splits DS-NP (D_{train}) into three smaller training sets (D_{train1}: 7073, D_{train2}/D_{train3}: 7181 each)—yielded three models (M_1, M_2 , M_3). Model M_1 emerged as the strongest, achieving 98.43% accuracy with lower test loss (0.0537), outperforming both the baseline and other DEV models (M_2: 98.22%, M_3: 98.15%). A significant reduction in FP was achieved by Model M_1 (35 vs. baseline's 181), indicating superior Negative detection. This suggests that smaller, curated datasets (like D_{train2}) may mitigate overfitting and improve generalization. The CNN baseline's balanced misclassification behavior is demonstrated through its FN (28 vs. baseline's 2). Regarding the precision, recall, and F1 comparison, Model M_1 exhibits the most balanced and highest scores across all metrics compared to Model M_2 and Model M_3. This suggests that Model M_1 provides the most robust and reliable performance among the three models.

The DEV method achieved remarkable performance improvements for DS-NP dataset, increasing accuracy by 3.58% (95.42% ⇒ 99%) while adding only 9 minutes and 40 seconds to execution time (3h55m37s ⇒ 4h5m17s). Error analysis shows exceptional optimization, with FP reduced by 83.4% (181 ⇒ 30) and FN increasing by 500% (2 ⇒ 12) due to the model's more conservative classification approach. The confusion matrix reveals outstanding TP detection (8.3% improvement, 1819 ⇒ 1970) with nearly perfect TN retention (1998 ⇒ 1988, -0.5%). These results demonstrate DEV's ability to push model performance to near-perfect levels while maintaining reasonable computational efficiency.

4) Performance evaluation for DS-BNB dataset: The baseline CNN model, trained on the full DS-BNB dataset (65.92K images), achieved a test accuracy of 89%, as detailed in Table IV, with a notable test loss of 0.281, indicating potential overfitting (train loss: 0.058, validation loss: 0.336). The confusion matrix revealed 312 FP (Bio images misclassified as $Non\text{-}Bio$) and 128 FN ($Non\text{-}Bio$ misclassified as Bio), suggesting challenges in distinguishing high-quality Non & Biodegradable material images.

In contrast, the DEV method—which splits DS-BNB (D_{train}) into three smaller training sets (D_{train1}: 19774, D_{train2}/D_{train3}: 20074 each)—yielded three models (M_1, M_2 , M_3). Model M_1 emerged as the strongest, achieving 89.6% accuracy with lower test loss (0.279), outperforming both the baseline and other DEV models (M_2: 83.75%, M_3: 86.83%). A significant reduction in FP was achieved by Model M_1 (192 vs. baseline's 312). Regarding the precision, recall, and F1 comparison, Model M_1 exhibits the most balanced and highest scores across all metrics compared to Model M_2 and Model M_3. This suggests that Model M_1 provides the most robust and reliable performance among the three models.

The DEV method delivered significant performance improvements for DS-BNB dataset, boosting accuracy by 4% (89% ⇒ 93%) with only 12-minute, 8-second increase in execution time (7h27m ⇒ 7h39m8s). The approach demonstrated particularly strong error reduction, with FP decreasing by 48.1% (312 ⇒ 162) while maintaining comparable FN rates (6.25%, 128 ⇒ 120). The confusion matrix reveals excellent TP improvement (8.9% increase, 1688 ⇒ 1838) with stable TN performance (1872 ⇒ 1880, 0.4%). These results highlight DEV's ability to enhance model precision while maintaining recall.

Finally, the DEV method's performance gains stem from four synergistic mechanisms. First, dataset partitioning enables specialized feature learning, where each model captures distinct data aspects, reducing collective bias. Second, the majority voting mechanism cancels individual model errors through consensus. Third, aggregated predictions leverage complementary model strengths, enhancing robustness to outliers and edge cases. Finally, the framework achieves this through a favorable balance between efficiency and accuracy, a small increase in execution time results in a disproportionate increase in accuracy, demonstrating a scalable improvement.

V. Conclusion

The Divided Ensemble Voting (DEV) methodology introduced in this research establishes a novel ensemble learning architecture aimed at augmenting the efficacy of image classification systems. Central to this framework is the strategic division of the training data into unique partitions, enabling the development of specialized convolutional neural networks tailored to distinct feature domains. The consensus-driven integration of predictions via a majority voting protocol yields a robust final classification. Empirical validation across multiple datasets confirmed a marked elevation in model accuracy, with observed gains ranging from 3.58% to 7.55%. The framework also demonstrated a profound capacity for error reduction, achieving a decrease in FP instances of up to 83.4%. These enhancements are attributed to the method's inherent resistance to overfitting, its ability to neutralize singular model biases through diversified training regimes, and its superior generalization powered by collective decision-making. Furthermore, DEV exhibited exceptional proficiency in addressing class distribution imbalances, reflected by a 16.5% increase in TP identification and F1-scores attaining 0.99. A critical advantage of this architecture is its computational efficiency, imposing a negligible runtime overhead of less than 5%, thereby affirming its practicality for operational environments. The combination of heightened precision and operational efficiency positions the DEV framework as a transformative solution for applications demanding high reliability.

REFERENCES

[1] N. Manakitsa, G. S. Maraslidis, L. Moysis, G. F. Fragulis, "A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision", *Technologies*, vol. 12, no. 2, p. 15, 2024.

[2] R. Archana, P. E. Jeevaraj, "Deep learning models for digital image processing: a review", *Artificial Intelligence Review*, vol. 57, no. 1, p. 11, 2024.

[3] S. A. El-aal, N. I Ghali, "A Proposed Recognition System For Alzheimer's Disease Based On Deep Learning And Optimization Algorithms", *Journal of Southwest Jiaotong University* , vol. 56, no. 5. 2021.

[4] A. Hassan, M. Refaat, A. M. Hemeida, "Image classification based deep learning: A Review", *Aswan University Journal of Sciences and Technology*, vol. 2, no. 1, pp. 11-35, 2022.

[5] Z. H. Arif, M. A. Mahmoud, K. H. Abdulkareem, M. A. Mohammed, M. N. Al-Mhiqani, A. A. Mutlag, R.Damaševičius, "Comprehensive review of machine learning (ML) in image defogging: Taxonomy of concepts, scenes, feature extraction, and classification techniques", *IET Image Processing*, 16(2), pp.289-310, 2022.

[6] B. F. Azevedo, A. M. A. Rocha, A .I. Pereira, "Hybrid approaches to optimization and machine learning methods: a systematic literature review, *Machine Learning*, vol. 113, no. 7, pp.4055-4097, 2024.

[7] I .U .Khan, S .Azam, S .Montaha, A .Al Mahmud, A .R .H .Rafid, M .Z .Hasan, M .Jonkman, "An effective approach to address processing time and computational complexity employing modified CCT for lung disease classification". *Intelligent Systems with Applications*, vol. 16, p.200147, 2022.

[8] G. Menghani, "Efficient deep learning: A survey on making deep learning models smaller, faster, and better" *ACM Computing Surveys*, vol. 55, no. 12, pp.1-37, 2023.

[9] Y. Yu, "Deep learning approaches for image classification", *In Proc. of the 2022 6th international conference on electronic information technology and computer engineering*, pp. 1494-1498, 2022.

[10] S. A. El-aal, R. S. El-Sayed, A. A. Alsulaiman, M. A. Razek, "Using Deep Learning on Retinal Images to Classify the Severity of Diabetic Retinopathy", *International Journal of Advanced Computer Science & Applications*, vol. 15, no.7, pp. 346-355, 2024.

[11] G. Vrbančič, V. Podgorelec, "Transfer learning with adaptive fine-tuning", *IEEE Access*, vol. 8, pp.196197-196211, 2020.

[12] M. Gholizade, H. Soltanizadeh, M. Rahmanimanesh, S. S. Sana, "A review of recent advances and strategies in transfer learning", *International Journal of System Assurance Engineering and Management*, pp.1-40, 2025.

[13] F. Maleki, K. Ovens, R. Gupta, C. Reinhold, A. Spatz, R. Forghani, "Generalizability of machine learning models: quantitative evaluation of three methodological pitfalls", *Radiology: Artificial Intelligence*, vol. 5, no. 1, p.e220028., 2022.

[14] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, T. Ganslandt, "Transfer learning for medical image classification: a literature review", *BMC medical imaging*, vol. 22, no. 1, p.69, 2022.

[15] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition. *In Proc. of the IEEE conference on computer vision and pattern recognition* ,pp. 770-778, 2016.

[16] E .S. Sabry, S. S. Elagooz, F. E. A. El-Samie, N. A. El-Bahnasawy, G. M. El-Banby, R. A. Ramadan, "Evaluation of feature extraction methods for different types of images", *Journal of Optics*, vol. 52, no. 2, pp.716-741, 2023.

[17] Y. C. Huang, C. W. Ho, W. R. Chou, M. Chen, "A framework to predict second primary lung cancer patients by using ensemble models. *Annals of operations research*, vol. 348, no. 1, pp.373-397, 2025.

[18] N. Nakata, T. Siina, "Ensemble learning of multiple models using deep learning for multiclass classification of ultrasound images of hepatic masses", *Bioengineering*, vol. 10, no. 1, p.69, 2023.

[19] Url = "https://zenodo.org/records/5528418#.YpdlS2hBzDd".

[20] Karras, T., Laine, S. and Aila, T., 2019. "A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4401-4410).

[21] Yasin, E.T. and Koklu, M., 2023, April. Classification of organic and recyclable waste based on feature extraction and machine learning algorithms. In Proceedings of the International Conference on Intelligent Systems and New Applications (ICISNA'23).

[22] N. Nnamoko,J. Barrowclough,J. Procter, "Waste Classification Dataset", *Mendeley Data*, 2022, https://doi.org/10.17632/n3gtgm9jxj.2.

[23] O. Omoebamije, Concrete & Pavement Crack Dataset, *Mendeley Data*, 2023, https://doi.org/10.17632/429vzbgmbx.1.

[24] M. Srimathi, M. Srivani, K. P. Peeyush, "Deep learning based waste material classification", *In Proc. of 2023 IEEE Silchar Subsection Conference (SILCON)*, pp. 1-6, 2023.

[25] https://www.kaggle.com/datasets/rayhanzamzamy/non-and-biodegradable-waste-dataset.

A Spatiotemporal Forex Trading System Based on a Hybrid Model GAT-LSTM: Forecasting Forex Price Directions

Nabil MABROUK, Marouane CHIHAB, Younes CHIHAB
Computer Sciences Laboratory, Ibn Toufail University, Kenitra, Morocco

***Abstract*—Due to the high volatility and complex interdependencies within financial markets, predicting Forex prices becomes a difficult challenge for investors. Furthermore, the traditional trading models struggle to capture those relationships. To address this issue, we introduced a spatiotemporal Forex trading system, GAT-LSTM-based; it is a hybrid approach that combines Graph Attention Network (GAT) with a Long Short-Term Memory (LSTM) network. The GAT component helps to capture spatial dependencies between currencies by constructing a directed graph containing 28 currency pairs alongside commodity stock and US stocks. The strength of the GAT component lies in its ability to dynamically adjust and recalculate the weights of edges over time, which helps our proposed system to adapt to macroeconomic changes, news events, and financial factors that can impact the Forex market status. The LSTM component deals with the nature of time series datasets. It learns temporal interdependencies, allowing our system to detect repeated long-term patterns over time. Experimental results proved that the suggested hybrid model, GAT-LSTM, surpasses both LSTM and GAT separately. By combining both elements and leveraging simultaneously the strength of dynamically modelling spatial dependencies, and the strength of learning long-term temporal patterns, our suggested system became more accurate in forecasting Forex price directions, showing promising results and high accuracy during the validation phase.**

Keywords*—Forex trading; hybrid deep learning model; Graph Attention Network (GAT); Long Short-Term Memory (LSTM); spatiotemporal forecasting

I. Introduction

The Forex, which is a short abbreviation of foreign exchange market, is known for its high volatility and strong price fluctuation caused by the influence of economic news, investor sentiment, and global events [1], [2]. Thanks to Forex advantages, including high liquidity, low transaction costs, 24-hour availability/operations, and online trading possibilities [3], investors (companies, individuals, and institutions) consider Forex as a profitable market that offers a good opportunity to achieve their goals and specific financial objectives utilizing exchange rate volatility [4], [5]. Most currencies are traded against the USD, and they are traded in pairs [6]. Each pair is composed of base and quoted currency [7]. Traders can exchange currencies at any time because the Forex market operates continuously for 24 hours a day, five days a week, which leads to elevated liquidity and a substantial trading volume [8], [9]. The historical data of Forex transactions has become the fuel of research.

In the past, numerous studies have been conducted on predicting Forex prices. Investors have deployed and adapted traditional trading strategies and other tools that they found helpful in their previous operations to suit their trading styles [10]. Three main techniques of forecasting Forex prices for time series are adopted by traders and investors: The first is fundamental analysis, which involves examining economic factors, political characteristics, and news events that may influence the Forex market status [11], [12]. The second is technical analysis (TA), which is a trading method based on historical data and mathematical calculations. To get an idea about future price variations, technical analysts look for past behaviors and patterns in historical data, using technical indicators and charts. The third type of analysis is sentiment analysis. Forex market fluctuations are primarily influenced by emotions and sentiments [13]. It studies the general mood of the market and analyzes the general attitude of traders toward the Forex market. Forex prices fluctuate because of investors' emotions, whether they are optimistic or pessimistic about the market. It is the primary factor for their decision-making to buy or sell. Trading techniques follow technological advancements.

Traders use computational algorithms, specifically deep learning, data mining, and machine learning, to create their trading strategies [14]. However, simple machine learning and deep learning algorithms struggle to forecast Forex prices accurately because of the social, economic, and political interdependencies.

In this study, we propose a hybrid model to leverage the dependencies between Forex currency pairs, the commodity stocks, and the US-stocks. Our spatiotemporal system can automatically forecast and identify future price directions, trends, and cycles, leveraging the time series nature of the Forex dataset. The spatial part of our model will deal with horizontal dependencies between currency pairs and stocks, forming a Graph to overcome the issue of economic, social, and political factors; each node of this Graph is enriched with valuable features such as technical indicators and trading rules to improve the prediction accuracy. Besides forming a Graphical representation, and thanks to the GAT component, which can update the strength of those connections between nodes, our model is able to adapt to global changes. The temporal part is to deal with the vertical dependencies between Forex transactions by learning and capturing long-term patterns. Our solution can perform the optimal trade decision, including buy, hold, or sell signals.

II. Related Works

In the last decade, forecasting currency exchange rates has been the center of much research; many techniques and

methodologies have been adopted in this field of research. Technological development helped investors trade at any time and from any place around the world through electronic platforms. [11], leading to the accumulation of a large set of historical data; which motivated researchers to conduct studies on the resolution of financial market problems based on computational algorithms [15]. In [16], Neely and Weller's research examines the manner in which technical analysis is employed in the foreign currency market, highlighting its widespread acceptability and superiority compared to fundamental analysis, particularly in short-term trading scenarios. Initially observed by traders in the London market, their findings have been validated across various market conditions, underscoring the enduring significance of technical analysis in Forex trading. In [17], they provide significant insights into the fundamental concepts of technical analysis, emphasizing its reliance on historical stock price and volume data to discern market patterns. Unlike fundamental analysis, which evaluates a stock's intrinsic worth, technical analysis emphasizes chart patterns and statistical indicators to forecast future market trends, offering a distinct perspective on the evaluation of equity assets. In [18], they conducted research on using SVM for market trend prediction; the results showed an out-performance of SVM with Expert Advisor, versus a simple Expert Advisor. They used an SVM (support vector machine) to forecast trends, "UP trend" or "DOWN trend". In [19], they showed that KNN performed better using PCA. KNN can reduce redundant and noisy information, leading to accurate predictions of the stock and Forex markets.

In [20], Zhelev and Avresky assert that the referenced literature on deep learning helps as a fundamental basis for handling the complex problem of Forex price prediction. In recent years, deep learning has shown notable progress in resolving Forex trading tasks and other domains [21]. Deep Neural Network algorithms have been employed to forecast time series data and quantitative trading [22], [23], [24]. Deep neural network algorithms have shown improved performance and accuracy, producing better returns in the finance sector [25], [26]. According to Galeshchuk [27], an ANN model with a single hidden layer can outperform the time series model in forecasting the price, especially for short-term prediction of the market. Hu et al. [28] applied NLP techniques in news to select only authentic content and influence trend. They proposed a hybrid LSTM model with an Attention mechanism for the prediction task. In [29], they confirmed that ANN performs better than DNN and SVM in time series data. In [30], the authors indicated that, while methodologies utilizing neural networks demonstrate significant potential, they are frequently evaluated solely on the basis of their prediction errors. However, it is equally imperative to assess their impact on a real-world trading system, where the decision to invest in the market is as critical as the decision not to invest.

In [31], they proposed time series data mining based on GNN and transformer to tackle the problem of modeling complex dependencies, using an adaptive adjacency matrix. They showed that the model can dynamically calculate the relationship between variables and capture local dependencies. For a multivariate time series forecast, in [32], they developed an architecture that merges a spatial-temporal graph neural network with a filtering module. They forecast future sales from an artificial time series sales volume dataset. The suggested spatial-temporal model showed excellent performance compared to baseline approaches without graphical architecture. In [33], they used a graph-awareness-based GAT-GAN Generative Adversarial Network to learn temporal dependencies that capture spatial relationships and generate long-time series data. In [34], they combined long-short-term memory (LSTM) with graph attention networks (GAT), creating a spatiotemporal model to accurately predict the water quality of the Pearl River Basin. They confirmed that their model showed robust performance and succeeded in jointly capturing spatial and temporal interdependencies. In [35], the authors examined the influence of human movement on the transmission of COVID-19. It establishes a mobility network that delineates travel patterns among Brazilian cities and incorporates time series data. Two Graph Convolutional Network (GCN) neural networks are employed to forecast time series for each city, considering adjacent cities. The study assesses LSTM and Prophet models, concluding that GCN models produce more consistent results than the assessed models. The Prophet model exhibits a greater maximum RMSE value.

III. Our Proposed Solution

A. Methodology and Problem Definition

The global economy is interconnected, a phenomenon known as internationalization, causing the interdependency of multiple financial sectors. However, Forex refers to the exchange of currencies. This fact implies the existence of dependencies between currencies. We assume that each pair of currencies is related to other pairs. Forming a directed graph G = (V, E) [36], where the set of nodes V represents all currency pairs, linked by an edge set E. these connections and the structure of the graph G are represented by a matrix adjacency A of two dimensions (D x D), where D is the number of nodes (in our case it is currency pair); each element A_{ij} represents the existence of an edge between two nodes V_i and V_j [see Eq. (1)]:

$$A_{ij} = \begin{cases} 1, & \text{if an edge exists between } V_i \text{ and } V_j \\ 0, & \text{Otherwise} \end{cases} \quad (1)$$

The Graph Neural Network (GNN) uses an adjacency matrix to capture spatial dependencies between currency pairs and to learn deep complex patterns. Providing the necessary information for accurate predictions. The GAT-LSTM model is trained using a supervised learning approach, which consists of minimizing the disparity between the forecasted and actual trends in historical data. The backpropagation during the training process allows the model to learn the optimal weights for both the GAT and LSTM components.

B. Data Description

1) Data collection: This study aims to predict the variation of prices in the Forex market. For that, we obtained the necessary dataset from various reliable financial platforms such as ForexSB, FXStreet, DailyForex, etc., which offer valuable historical Forex data for time series analysis.

To construct comprehensive data for our model, we considered the 28 most common currency pairs. They can be

classified into two general classes: the first is major currencies (GBPUSD, EURUSD, USDCHF, USDCAD, and USDJPY), and the second class is the minor currencies (AUDCAD, NZDJPY, NZDUSD, CADCHF, NZDCAD, GBPNZD, NZDCHF, GBPCHF, GBPJPY, GBPAUD, AUDNZD, EURNZD, EURJPY, AUDCHF, EURGBP, EURCHF, EURCAD, EURAUD, CHFJPY, CADJPY, AUDUSD, GBPCAD, AUDJPY).

The objective is to forecast future price directions for the five major currency pairs. In addition to the 28 currency pairs, we used six additional stock data, such as commodity stocks and US stocks, to enhance the predictive performance of our algorithm. The collected dataset covers the period from January 2009 until June 2025, spanning 5006 days, resulting in approximately 170,204 daily observations (5006 days x 34 stocks). This dataset provides a detailed view of the exchange rates for the currencies, presented by five features:(Open, High, Low, Close, Volume) for each stock. Where Open is the price at the beginning, High is the highest price, Low is the lowest price, Close is the final price, and Volume is the total amount of currency exchanged. Fig. 1 illustrates the scaled closing prices of the five major currency pairs.

2) Data preprocessing and feature engineering: In the data pre-processing phase, we handled the missing values to avoid errors in prediction models by ensuring that there were no missing values in our time series dataset. In addition to primary features, we generated two other variables, "return" and "direction" features, that indicate whether the prices will increase or decrease in the following sequence [see Eq. (2)]:

$$Direction = \begin{cases} 1, & \text{if Price } _{t+1} > \text{ Price } _{t} \\ 0, & \text{Otherwise} \end{cases} \quad (2)$$

To improve the performance of our model, we created additional features that are generated using mathematical calculations. Those features are called technical indicators, such as:

- The Bollinger Band (BB),
- Moving Polynomial Trending (MPT)
- The Relative Strength Index (RSI),
- The Commodity Channel Index (CCI),
- The Weighted, Exponential, Simple, and Convergence Divergence Moving Average (WMA, EMA, SMA, and MACD),
- The average directional index (ADX),
- The Rate-of-Change (ROC)

Technical indicators can be incorporated as additional features, enriching our dataset with additional information that helps to improve the accuracy of the prediction. We finally form a comprehensive spatiotemporal 3D matrix for model training and evaluation (5006 days x 34 stocks x 81 features).

3) Feature matrix X: Let X be a feature matrix of dimension (N x P), where N represents the total number of currency pairs and P is the total number of features of each currency pair. The features matrix $X \in \mathbb{R}^{N \times P}$ is defined by the following Eq. (3):

$$X^k = \begin{bmatrix} x_{11}^k & \cdots & x_{1p}^k \\ \vdots & \ddots & \vdots \\ x_{n1}^k & \cdots & x_{np}^k \end{bmatrix} \quad (3)$$

Our dataset is a 3D matrix, where X^k represents a feature matrix at row k.

4) Feature scaling: During the training process, features with larger magnitudes dominate, preventing the model from learning. To optimize the efficacy of our model and guarantee that all features contribute equally to the training process, we have scaled the data using the Min-Max scaler in the range [0, 1]. See Eq. (4):

$$X_{sc} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \quad (4)$$

where,

Xsc: the scaled value of the data point.

X: the initial data point's value.

Xmin: the minimal value of the data point.

Xmax: the maximum value of the data point.

5) The spatial representation of currency pairs (Graph): We assume that the currency pairs form a graph. Each currency pair is considered a node; the relationships between those nodes (edges) are determined based on the calculation of correlation coefficients, forming a weighted graph. Those weights are dynamic and change during training, enabling our model to self-adjust with global factors such as economic cycles, news events, and industry trends. Initially, we created a fully connected graph: Every currency pair is connected to every other currency pair. Fig. 2 illustrates the graph representation of currency pairs.

6) Sequence creation: Our objective is to improve the predictive capabilities of GAT and LSTM by effectively leveraging their combined power. We reconstructed the resulting data into sequences by implementing a sliding-window approach.

Let $X = \{x_1, x_2, ..., x_T\}$ be a time series dataset, where x_t represents the data point at time t, and T is the length of the initial dataset. The i^{th} sequence S_i of length L is defined in Eq. (5):

$$S_i = \{x_i, x_{i+1}, \ldots, x_{i+L-1}\} \quad (5)$$

For example, a window of five days length means that to forecast the price direction Yt of a currency pair, we feed our model with the data that correspond to time step t, along with the data of the four prior time steps $[x_{t-3},\ x_{t-3},\ x_{t-2},\ x_{t-1},\ x_t]$. This approach helps the LSTM component to better detect temporal dependencies. Furthermore, for better optimization of the LSTM layer, we have trained our model using various sequence lengths between 1 and 25.

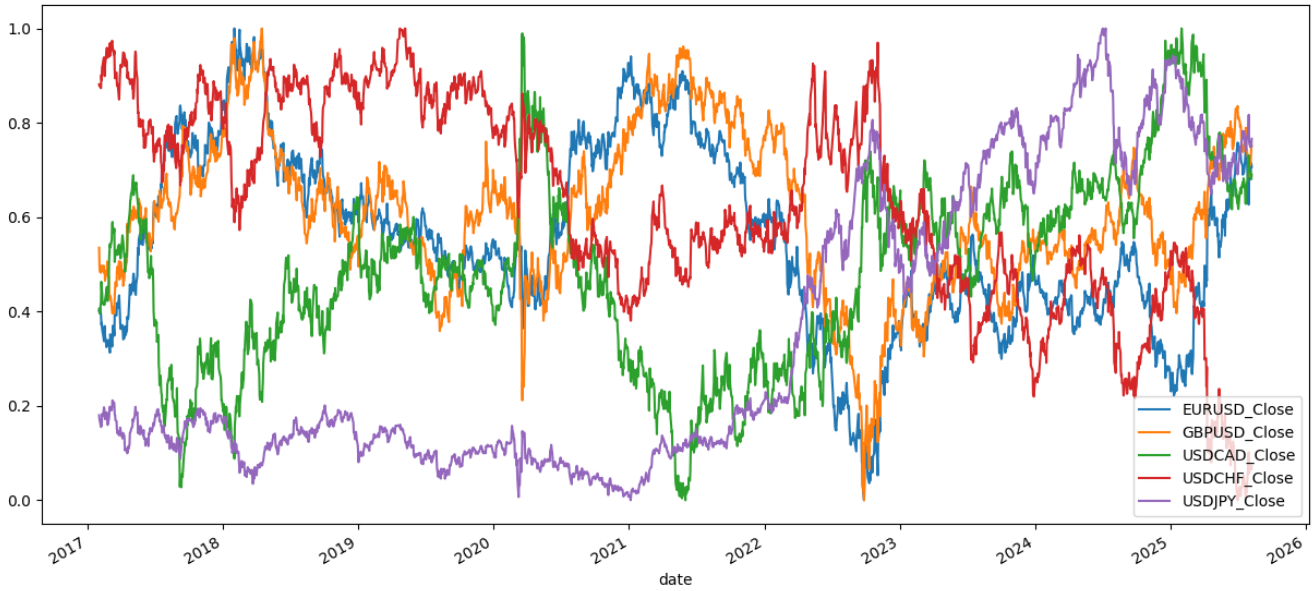


Fig. 1. The scaled closing price of the five major currency pairs.

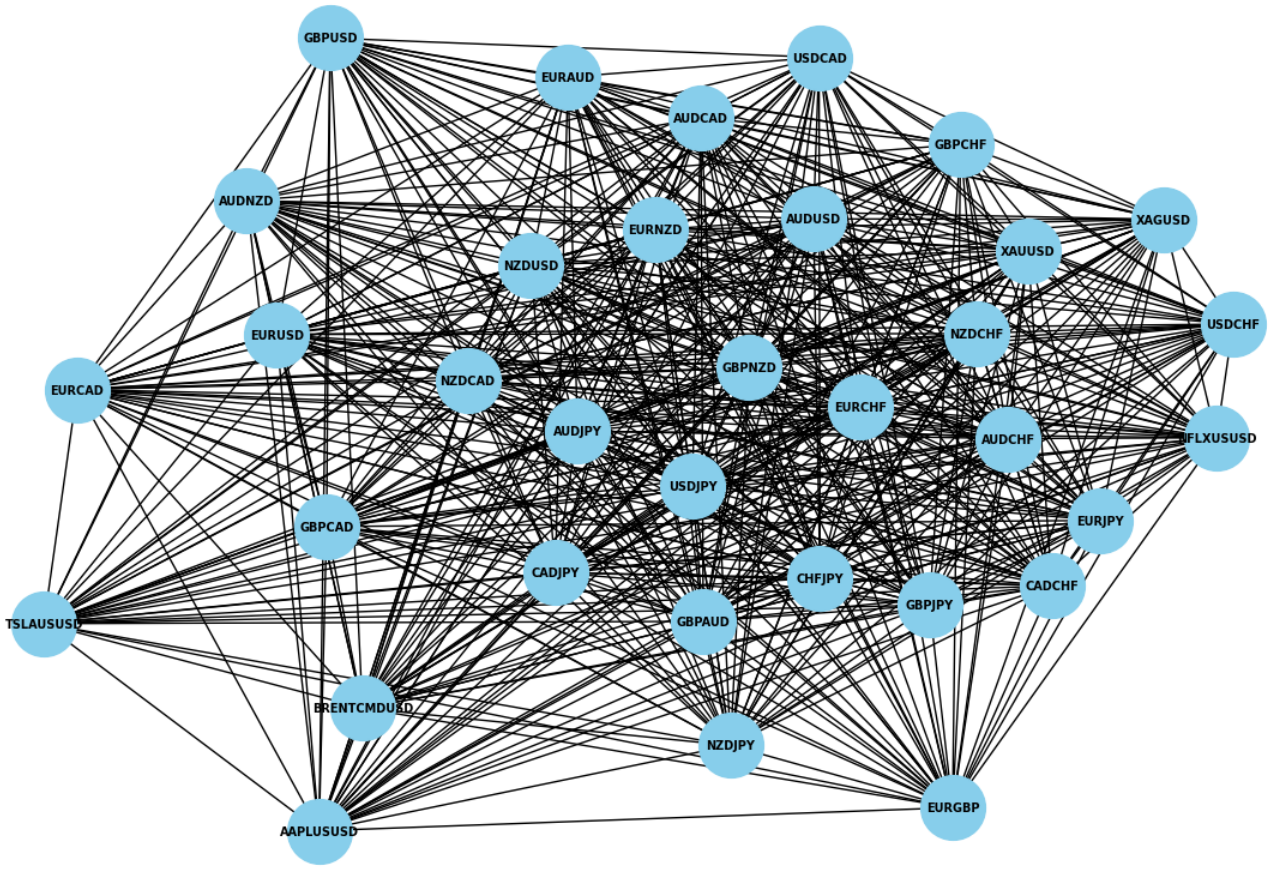


Fig. 2. A complete graph with 28 currency pairs and six stocks.

7) Split data into train and validation: To ensure a realistic and rigorous evaluation, we implemented a strict chronological data split to mimic real-world prediction. The dataset was divided temporally as follows:

- Training set (70%): January 1, 2009, to December 31, 2019.
- Validation set (10%): January 1, 2020, to December 31, 2021.
- Test set (20%): January 1, 2022, to May 30, 2025.

Crucially, we employed a strict procedure to prevent data leakage. All preprocessing steps were calculated solely from the training data, and those transformations were then applied to the subsequent validation and test sets. By maintaining this chronological order and separation, we were able to accurately assess the capacity of the model to generalize to future unseen data.

C. Model Architecture

In this study, we implemented two independent models, the first is Graph Attention (GAT) Network and the second is Long Short-Term Memory (LSTM) Network, to forecast Forex trend direction. We compared those models with a hybrid model, GAT-LSTM, which is a spatiotemporal model that combines two components to take advantage of the strengths of both architectures in dealing with spatial and temporal interdependencies simultaneously.

1) GAT (Graph Attention Networks): Like any Graph Neural Network (GNN) architecture, the Graph Attention Network (GAT) is designed to analyze and understand graph-structured data; in addition, it implements the attention mechanisms [37]. GAT assigns attention weights to neighboring nodes. Updates the node representation according to the importance of the neighbor's information and refines the relationships between individual pairs of currencies. This aids in generating a dynamic, weighted graph, which subsequently helps manage fluctuations in the Forex market. The following Eq. (6) mathematically defines the attention mechanism:

$$\alpha_{\mathrm{ij}} = \mathrm{softmax}_{\mathrm{j}} \left(a \left(W * h_{\mathrm{i}}, W * h_{\mathrm{j}} \right) \right) \tag{6}$$

Let α_{ij} be the attention coefficient between currency pair i and currency pair j, where h_i and h_j denote the characteristic vectors of currency pair i and currency pair j, respectively; W is a learnable weight matrix, while a is an attention function. The attention mechanism dynamically adjusts the importance of each edge based on the learned attention weights. By leveraging attention mechanisms, GAT can effectively overcome the issue of stock variations caused by various factors.

2) The Recurrent Neural Networks and the Gated Recurrent Unit: To address sequential data challenges, including time series and text, researchers have created the Recurrent Neural Network (RNN) [38], [39], [40], a specialized neural network that incorporates a memory mechanism to retain information from prior inputs. Feedback links in their design enable the retention of knowledge across temporal stages. Recurrent Neural Networks (RNNs) excel in sequence modeling problems, where the temporal context of data points is essential for precise prediction and comprehension.

3) LSTM (Long Short-Term Memory): Long Short-Term Memory (LSTM) was published by [41], [42], [43], [44]. It is an architecture of Recurrent Neural Networks (RNNs) that can help address the problem of vanishing gradients faced by standard RNNs. LSTM can effectively bridge time lags exceeding 1,000 discrete time steps. This capability makes them suitable for tasks where long-range temporal dependencies are critical.

The idea behind using an LSTM layer in our hybrid model is to learn and recognize sequential patterns that characterize time series datasets and capture temporal dependencies of Forex prices. We have adopted a many-to-many architecture to forecast the following trend of the five major currency pairs.

4) Hybrid model GAT-LSTM: In this study, we have proposed a hybrid model to effectively forecast Forex trends by capturing spatial dependencies between currency pairs and learning sequential patterns simultaneously. This model, GAT-LSTM, is a combination of two components, GAT and LSTM; see Fig. 3. By integrating the GAT and LSTM networks, we succeeded in creating a spatiotemporal model that takes advantage of the strengths of the first component, GAT, in dealing with spatial dependencies in the Forex by defining edges between currencies. GAT helps adjust the weights of edges while learning and training, which enables our model to adapt to economic changes. The second component, LSTM, is designed to handle the nature of time series datasets and detect temporal patterns that are repeated over time.

The model procedure begins with integrating multidimensional features into a predefined directed graph. The GAT module applies attention mechanisms, allowing each node to selectively absorb more information from neighboring nodes based on relevancy weights. This capability enables the model to handle diverse inputs from various sources, which results in spatially enhanced feature representations. Then, these spatial representations are reorganized into chronological patterns to be used by the LSTM module for capturing dynamic patterns in Forex prices, such as trends and cycles. The strength of LSTM lies in effectively modeling both short-term variability and long-term dependencies, which are critical to accurately forecasting dynamic currency prices.

The final step is the fully connected layer (the output layer). It is a linear layer that takes the temporal features extracted by LSTM and generates predictions for the trend direction. The role of this layer is to perform a classification task; the output is a prediction of the direction of the currency price (buy, hold, or sell) of the five major currency pairs; the number of units corresponds to the number of predicted classes.

D. Training Phase (Hyperparameter Optimization and Fine-Tuning)

For efficient training, we ran our model for 50 epochs and finetuned it by optimizing the hyperparameters during the learning phase; the batch size is around 24 and 32. And for loss function optimization, we have used the Adam optimizer with a learning rate around 0.001. Our algorithm attempts to resolve a classification issue by forecasting the Forex trend direction. We employed BCEWithLogitsLoss for the validation and training loss function, which is used for multilabel classification problems. BCEWithLogitsLoss combines the sigmoid

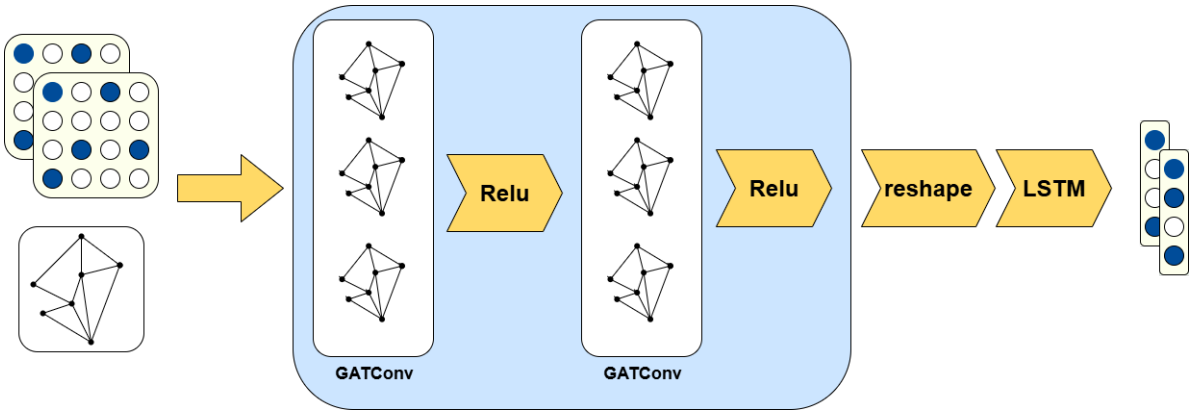


Fig. 3. The flowchart of the GAT-LSTM hybrid model for Forex price direction forecasting.

TABLE I. COMPARISON OF MODELS' PERFORMANCE (ACCURACY)

	LSTM	GAT	GAT-LSTM
EURUSD	0.6727	0.5206	0.8369
GBPUSD	0.6856	0.5026	0.8216
USDCAD	0.6134	0.4845	0.8267
USDCHF	0.6314	0.5077	0.8236
USDJPY	0.6418	0.5361	0.8287

and BCE loss into a single function; the formal definition is given in Eq. (7) as follows:

$$\text{Loss} = \text{mean}\left(BCE\left(\sigma\left(x_i\right), y_i\right)\right) \tag{7}$$

where, x_i and y_i represent the logits and true labels for the i^{th} element in the batch, respectively. The mean() function calculates the average loss over all samples in the batch. The loss can also be configured to return the sum of the losses instead of the mean, depending on specific needs.

IV. DISCUSSION AND RESULTS

A. LSTM Performance

The LSTM model is trained for 30 epochs using both architectures, with and without the window approach. In both methods, the model performed moderately in capturing temporal dependencies in currency price movements. The accuracy was between 61% and 67%, as shown in Table I, indicating that the LSTM model struggles to predict the general trend with specific turning points.

B. GAT Performance

We trained the GAT model for 50 epochs; its performance is slightly lower than that of LSTM. The accuracy did not surpass 50% which is not impressive, see Table I. These results could be caused by the graph structure, which considers only horizontal dependencies between nodes and ignores the vertical dependencies characterized by the sequential nature of stock price data. The weak correlation between nodes was not enough to provide significant advantages.

C. GAT-LSTM Performance

Compared to GAT and LSTM models, the hybrid model GAT-LSTM showed better performance in forecasting Forex trends as shown in Table I. We trained GAT-LSTM for 30 epochs, using different window size as described in Table II. After analyzing the training and validation loss graphics, as illustrated in Fig. 4 and Fig. 5, we found that the model is well-trained and fits the training data well, and it can be generalized for a larger time frame. We observed the lowest accuracy value (72%) at the lowest window size (1 day); while the highest accuracy value (83.7%) is observed at 20 days window length; the accuracy corresponds to the sequence length; When the window length is higher than 20 days, the algorithm becomes slower and less performant, the accuracy starts decreasing. We conclude that the optimal size of the sequence length is 20 days. The fact that the model's performance corresponds to the size of the window is because feeding the model with additional significant features helps identify patterns.

The high performance of LSTM-GAT is relevant to the fact that it is a hybrid model combining temporal modeling (LSTM) with relation modeling (GAT), which leads to better learning of Forex market dynamics. The LSTM is responsible for capturing individual Forex price patterns, while GAT captures the dependencies between currencies. That helps improve the predictions.

D. Generalization and Performance During Conflict

While our primary focus is on demonstrating the effectiveness of the GAT-LSTM model for Forex price direction forecasting, we recognized the need to account for real-world trading frictions like transaction costs (spreads) and slippage. To provide a more realistic perspective, we applied a moderate 10% reduction to backtested profits achieved during the period test, specifically the COVID-19 pandemic. Even with this adjustment, our system exhibits strong potential for profitability and resilience during crises.

1) Return on Investment (ROI): We evaluated the performance of our model using the same loss function, BCE-WithLogitsLoss, and calculating the accuracy of prediction as shown in Fig. 6 and Fig. 7. Furthermore, we back-tested our proposed system using financial metrics that consist of calcu-

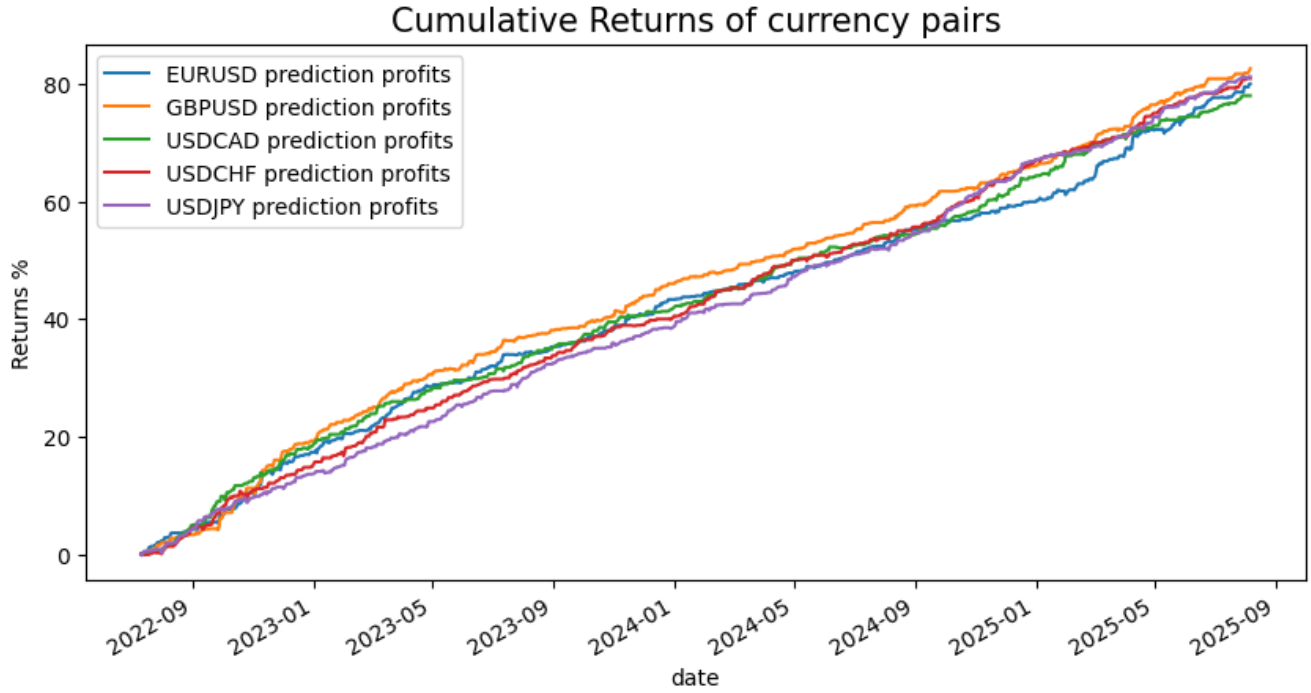


Fig. 4. GAT-LSTM training/validation loss using 20 days window size.

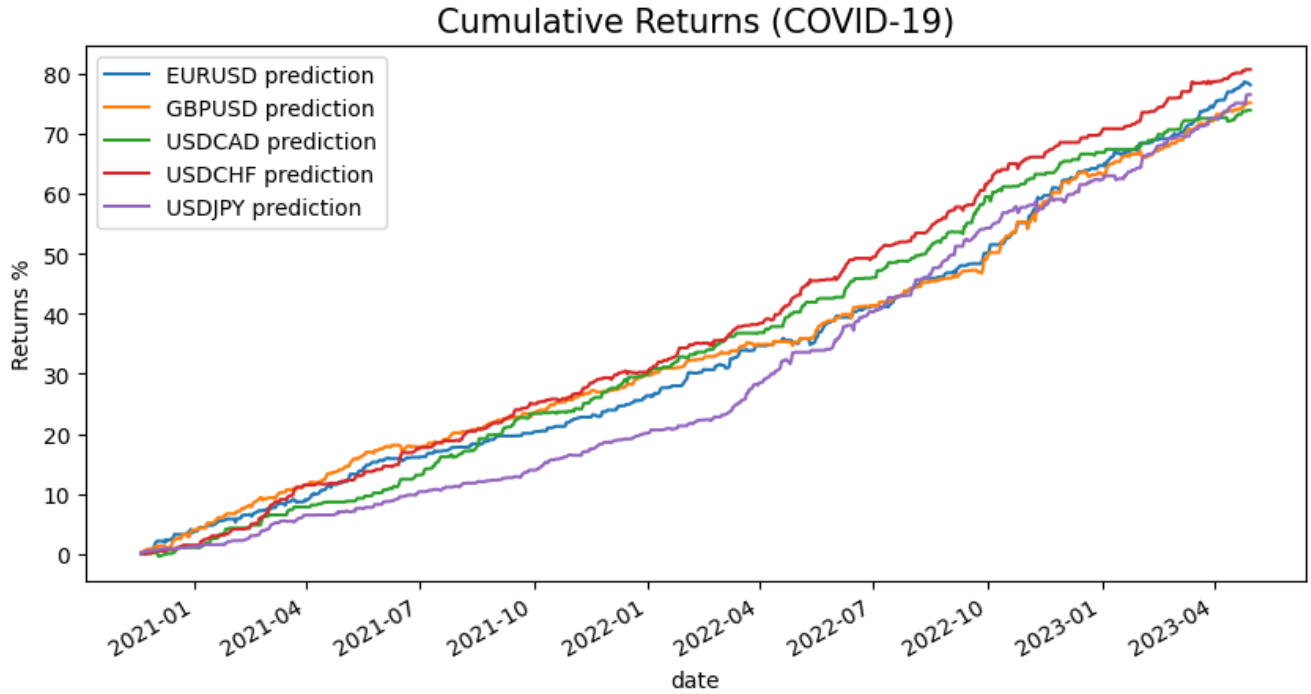


Fig. 5. GAT-LSTM training/validation accuracy using 20 days window size.

TABLE II. MODEL ACCURACY BY WINDOW-SIZE AND TIME FRAME

Currency pair	Window size					
	1day	5 days	10days	15days	20days	25days
EURUSD	0.7462	0.7686	0.8287	0.8165	0.8369	0.8217
GBPUSD	0.7339	0.7472	0.7910	0.7961	0.8216	0.8147
USDCAD	0.7309	0.7503	0.7910	0.8135	0.8267	0.8094
USDCHF	0.7523	0.7564	0.8114	0.8114	0.8236	0.8159
USDJPY	0.7288	0.7666	0.8063	0.8022	0.8287	0.8136

TABLE III. DIEBOLD-MARIANO TEST RESULTS (P-VALUES)

Model Pair	DM Test Statistic	p-value	Conclusion ($\alpha = 0.05$)
GAT-LSTM vs Naïve Bayes	8.12	3.15×10^{-13}	Significant (GAT-LSTM is better)
GAT-LSTM vs Logistic Regression	10.02	1.12×10^{-17}	Significant (GAT-LSTM is better)
GAT-LSTM vs Random Forest	7.55	2.50×10^{-11}	Significant (GAT-LSTM is better)
GAT-LSTM vs XGBoost	3.54	8.50×10^{-04}	Significant (GAT-LSTM is better)

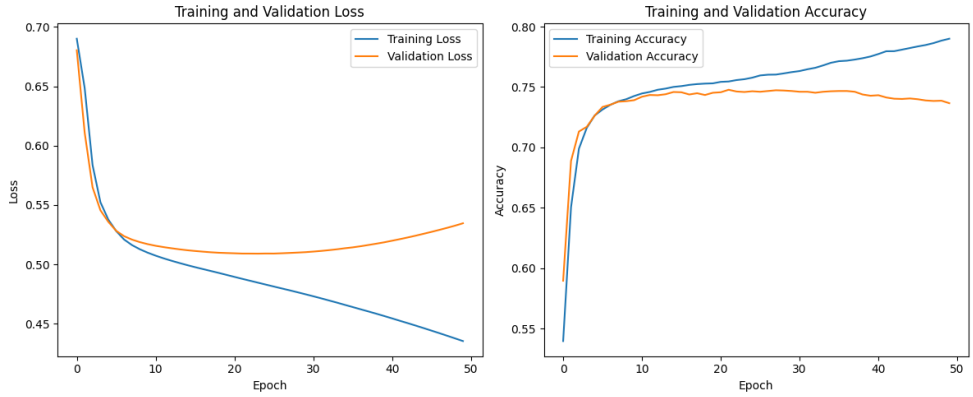


Fig. 6. The cumulative returns of the five major currency pairs during the COVID-19 pandemic.

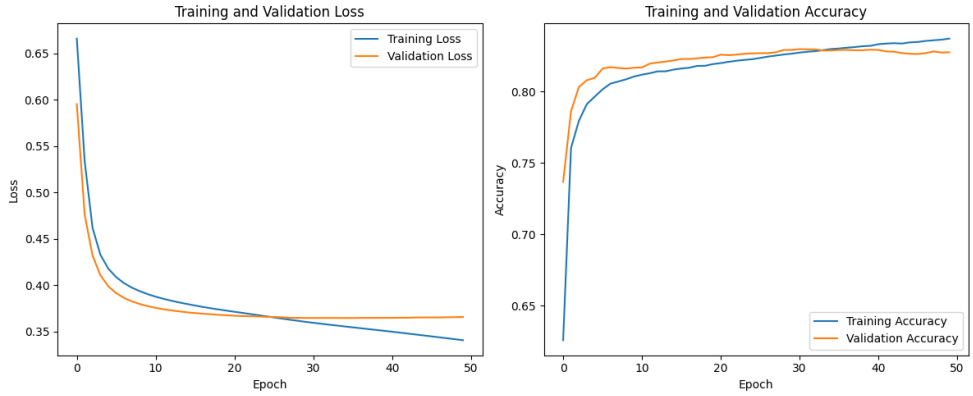


Fig. 7. The cumulative returns of the five major currency pairs during the period test.

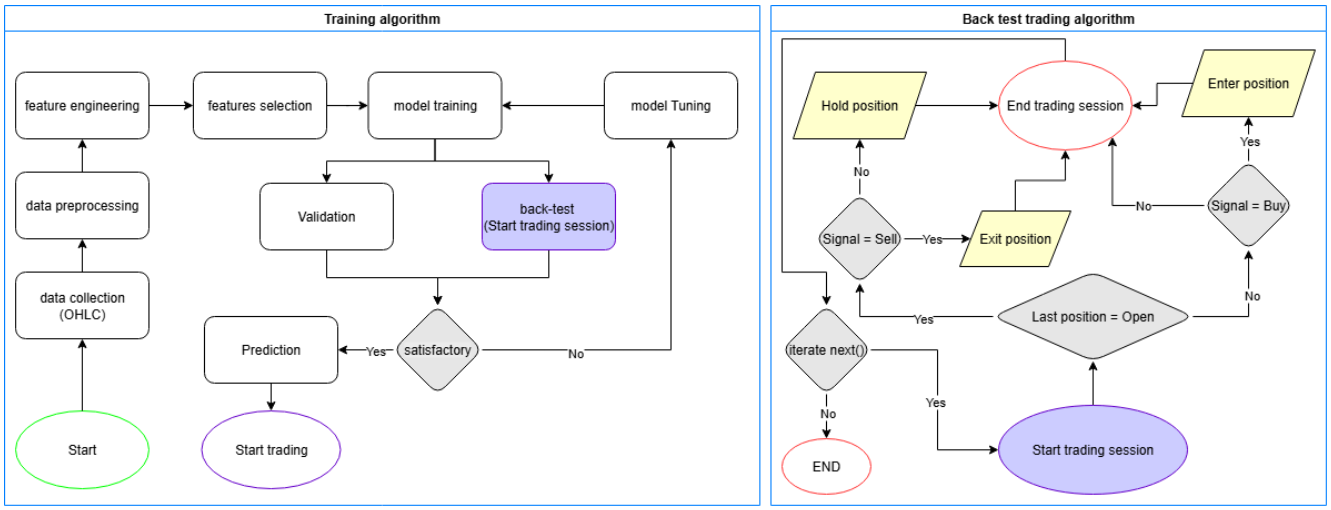


Fig. 8. Our investment strategy and the back-test flowchart.

lating the return profits of forecasted trends. Fig. 8 illustrates our trading strategy and the adopted back-test algorithm.

As illustrated in Fig. 6, our system successfully predicted the future direction of the prices, achieving an accuracy of 83.7% with a total profit of 79% compared to the ideal.

2) Performance during conflict: During Global economic crises, the Forex market can experience sharp swings caused by investor panic. The rising index of fear is leading to significant adverse ripple effects within the economy, resulting in a marked increase in market volatility. Every speculative attack on the Forex market frequently coincides with a crisis. We back-tested our model during the COVID-19 pandemic from March 2020 to May 2023.

As shown in the Fig. 7, which shows the cumulative returns of the five major currency pairs during the crisis period, our approach successfully forecasted the future direction of the prices, achieving an average accuracy of 81% corresponding to 73% of the total return profits compared to the ideal. Our suggested trading system can overcome obstacles during times of crisis thanks to the GAT component that can capture dependencies between stocks and currency pairs and anticipate global changes.

3) Statistical validation of hybrid model superiority: We used the Diebold-Mariano (DM) test to perform a rigorous statistical comparison of our hybrid GAT-LSTM model against traditional baseline models, assessing their predictive accuracy using a zero-one loss function. The DM test results, detailed in Table III, yielded p-values that were all significantly below the 0.05 significance level (α=0.05). This finding provides strong evidence that the performance gains of the GRU-CNN model are statistically significant, meaning that the improved accuracy is due to the model's superior ability to capture complex patterns in the financial time series data, rather than being a result of random chance. This analysis validates our choice of a deep learning architecture for this prediction task.

V. Conclusion

GAT-LSTM architecture integrates two learning mechanisms simultaneously, spatial and temporal, addressing the complexity and high fluctuations caused by macro-economic, political factors, and news events, etc., offering an effective solution for forecasting Forex prices. As a result, GAT-LSTM is well-equipped to adapt to the rapidly changing nature of Forex markets, providing traders with valuable insights for decision-making. However, currency prices do not depend on other currencies and stock markets only; many other factors can affect the exchange rate. We created and evaluated this solution for a typical period. It could stop being beneficial and lucrative in the future. Additional research must be done to enhance the graph structure, feature engineering, and model architecture for improved performance. Integrating temporal dynamics and inter-currency interactions improves the accuracy and robustness of forecasting models.

Data Availability Statement

Due to privacy and ethical restrictions, the data that support the findings of this study cannot be made publicly available. However, a simplified version of the codebase is available on request from the corresponding author.

References

[1] S. Ahmed, S.-U. Hassan, N. R. Aljohani, and R. Nawaz, "Flf-lstm: A novel prediction system using forex loss function," *Applied Soft Computing*, vol. 97, p. 106780, 2020.

[2] R. D. Huang and R. W. Masulis, "Fx spreads and dealer competition across the 24-hour trading day," *The Review of Financial Studies*, vol. 12, p. 61–93, 1999.

[3] D. C. Yıldırım, I. H. Toroslu, and U. Fiore, "Forecasting directional movement of forex data using lstm with technical and macroeconomic indicators," *Financial innovation*, vol. 7, p. 1, 2021.

[4] Y. Bilgin, S. M. Camgoz, M. B. Karan, and Y. Yildiz, *Understanding the investment behavior of individual investors: an empirical study on forex markets*. IGI Global Scientific Publishing, 2021.

[5] L. Puspitawati, "Information system for forex investment and their effects on investment growth in foreign currencies," *International Journal of Research and Applied Technology (INJURATECH)*, 2021.

[6] M. Ozturk, I. H. Toroslu, and G. Fidan, "Heuristic based trading system on forex data using technical indicator rules," *Applied Soft Computing*, vol. 43, p. 170–186, 2016.

[7] M. A. Junior, P. Appiahene, O. Appiah, and C. N. Bombie, "Forex market forecasting using machine learning: Systematic literature review and meta-analysis," *Journal of Big Data*, vol. 10, p. 9, 2023.

[8] H. Law, S. Aghabozorgi, S. Lim, Y. Teh, and T. Herawan, "Processing of kuala lumpur stock exchange resident on hadoop mapreduce," in *Proceedings of the International Conference on Data Science (ICDATA)*. Proceedings of the International Conference on Data Science (ICDATA), 2014, p. 1.

[9] G. McLeod, "Forex market size: A traders advantage," *DailyFX (January 23).*, 2014.

[10] Y. Pei, J. Cartlidge, A. Mandal, D. Gold, E. Marcilio, and R. Mazzon, "Cross-modal temporal fusion for financial market forecasting," *arXiv preprint arXiv:2504.13522*, 2025.

[11] L. Abednego, "Forex trading robot with technical and fundamental analysis," *Journal of Computers*, p. 1089–1097, 2018.

[12] K. Cao and H. You, "Fundamental analysis via machine learning," *Financial Analysts Journal*, vol. 80, p. 74–98, 2024.

[13] L. Abednego, C. E. Nugraheni, and A. Fedora, "Forex sentiment analysis with python," *International Journal of Advanced Research in Economics and Finance*, vol. 4, p. 46–55, 2022.

[14] N. MABROUK, M. CHIHAB, Z. HACHKAR, and Y. CHIHAB, "Intraday trading strategy based on gated recurrent unit and convolutional neural network: forecasting daily price direction," *International Journal of Advanced Computer Science and Applications*, vol. 13, 2022.

[15] E. Gerlein, M. Mcginnity, A. Belatreche, and S. Coleman, "Evaluating machine learning classification for financial trading: An empirical approach," *Expert Systems with Applications*, vol. 54, pp. 193–207, Jul. 2016.

[16] C. J. Neely and P. A. Weller, "Technical analysis in the foreign exchange market," *Handbook of exchange rates*, pp. 343–373, 2012.

[17] V. Drakopoulou, "A review of fundamental and technical stock analysis techniques," *Journal of Stock & Forex Trading*, vol. 5, 2016.

[18] T. N. T. Thu and V. D. Xuan, "Using support vector machine in forex predicting," in *2018 IEEE International Conference on Innovative Research and Development (ICIRD)*. 2018 IEEE International Conference on Innovative Research and Development (ICIRD), 2018, p. 1–5.

[19] L. Tang, H. Pan, and Y. Yao, "K-nearest neighbor regression with principal component analysis for financial time series prediction," in *Proceedings of the 2018 international conference on computing and artificial intelligence*. Proceedings of the 2018 international conference on computing and artificial intelligence, 2018, p. 127–131.

[20] S. Zhelev and D. R. Avresky, "Using lstm neural network for time series predictions in financial markets," in *2019 IEEE 18th international symposium on network computing and applications (NCA)*. 2019 IEEE 18th international symposium on network computing and applications (NCA), 2019, p. 1–5.

[21] L. Dongdong, S. Yuan, M. Li, and Y. Xiang, "An empirical study of machine learning algorithms for stock daily trading strategy," *Mathematical Problems in Engineering*, vol. 2019, pp. 1–30, Apr. 2019.

[22] D. Lv, Z. Huang, M. Li, and Y. Xiang, "Selection of the optimal trading model for stock investment in different industries," *PLOS ONE*, vol. 14, p. e0212137, Feb. 2019.

[23] B. Yue, J. Rao, and Y, "A deep learning framework for financial time series using stacked autoencoders and long short term memory," *PLoS ONE*, vol. 12, p. 1–24, 2017.

[24] R. Akita, A. Yoshihara, T. Matsubara, and K. Uehara, *Deep learning for stock prediction using numerical and textual information*. Okayama, Japan: IEEE, Jun. 2016.

[25] Z. Hu, Y. Zhao, and M. Khushi, "A survey of forex and stock price prediction using deep learning," *Applied System Innovation*, vol. 4, p. 9, Feb. 2021.

[26] H. Jia, "Deep learning algorithm-based financial prediction models," *Complexity*, vol. 2021, p. 5560886, 2021.

[27] S. Galeshchuk, "Neural networks performance in exchange rate prediction," *Neurocomputing*, vol. 172, p. 446–452, 2016.

[28] Z. Hu, W. Liu, J. Bian, X. Liu, and T.-Y. Liu, "Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction," in *Proceedings of the eleventh ACM international conference on web search and data mining*. Proceedings of the eleventh ACM international conference on web search and data mining, 2018, p. 261–269.

[29] S. Galeshchuk and S. Mukherjee, "Deep networks for predicting direction of change in foreign exchange rates," *Intelligent Systems in Accounting, Finance and Management*, vol. 24, p. 100–110, 2017.

[30] M. Fisichella and F. Garolla, "Can deep learning improve technical analysis of forex data to predict future price movements?" *Ieee Access*, vol. 9, pp. 153 083–153 101, 2021.

[31] J. Wang, "Multivariate time series forecasting and classification via gnn and transformer models," *Journal of Computer Technology and Software*, vol. 3, 2024.

[32] Y. Wang and T. Aste, "Network filtering of spatial-temporal gnn for multivariate time-series prediction," in *Proceedings of the Third ACM International Conference on AI in Finance*. Proceedings of the Third ACM International Conference on AI in Finance, 2022, p. 463–470.

[33] S. Iyer and T. T. Hou, "Gat-gan: A graph-attention-based time-series generative adversarial network," *arXiv preprint arXiv:2306.01999*, 2023.

[34] Y. Liu, H. Zheng, and J. Zhao, "Enhanced water quality prediction by lstm and graph attention network (l-gat): An analytical study of the pearl river basin," *Water Research X*, p. 100383, 2025.

[35] F. H. O. Duarte, G. J. Moreira, E. J. Luz, L. B. Santos, and V. L. Freitas, "Time series forecasting of covid-19 cases in brazil with gnn and mobility networks," in *Brazilian Conference on Intelligent Systems*. Springer, 2023, pp. 361–375.

[36] Y. Bai, G. Li, T. Lu, Y. Wu, W. Zhang, and Y. Feng, "Map matching based on seq2seq with topology information," *Applied Sciences*, vol. 13, p. 12920, 2023.

[37] A. G. Vrahatis, K. Lazaros, and S. Kotsiantis, "Graph attention networks: a comprehensive review of methods and applications," *Future Internet*, vol. 16, no. 9, p. 318, 2024.

[38] K. K. L. Wong, "The structure of neural network," 2024.

[39] F. M. Salem, *Recurrent neural networks*. Springer, 2022.

[40] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent neural networks: A comprehensive review of architectures, variants, and applications," *Information*, vol. 15, p. 517, 2024.

[41] hochreiter Sepp and schmidhuber jürgen, "Memory, long short-term," *Neural Computation*, vol. 9, p. 1735, 1997.

[42] S. Hochreiter and J. Schmidhuber, "Lstm can solve hard long time lag problems," *Advances in neural information processing systems*, vol. 9, 1996.

[43] F. A. Gers and E. Schmidhuber, "Lstm recurrent networks learn simple context-free and context-sensitive languages," *IEEE transactions on neural networks*, vol. 12, p. 1333–1340, 2001.

[44] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural computation*, vol. 12, p. 2451–2471, 2000.

Experimental Validation of an Adaptive Controller for a Mecanum-Wheel Robot with Unknown Center-of-Gravity Offset and Slope Inclination

Chawannat Chaichumporn[1], Supaluk Prapan[2], Nghia Thi Mai[3],
Md Abdus Samad Kamal[4], Iwanori Murakami[5], Kou Yamada[6]
Faculty of Engineering, Thai-Nichi Institute of Technology, Bangkok, 10250, Thailand[1,2]
Graduate School of Science and Technology, Gunma University, Kiryu, 376-8515, Japan[1,4,5,6]
Faculty of Electronics Engineering 1, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam[3]

***Abstract*—High-precision path tracking for a Four-Mecanum-Wheel Mobile Robot (FMWMR) is challenged by real-world factors such as payload-induced shifts in the center-of-gravity (CoG) and operation on inclined surfaces. These uncertainties introduce complex, coupled dynamic forces that degrade the performance of conventional controllers. This study addresses this problem with a Model Reference Adaptive Controller (MRAC), which learns and compensates for these unpredictable dynamic effects in real time. To ensure effective operation on physical hardware, the controller incorporates practical solutions for motor friction and control signal stability. The proposed approach is validated through implementation of the MRAC on a Rosmaster X3 robot. A performance comparison is made against a well-tuned Proportional-Integral-Derivative (PID) controller is conducted across twelve distinct scenarios. The results show that the adaptive controller reduced the position Root Mean Square Error (RMSE) by an average of 52.7% and the Integral of Time-weighted Absolute Error (ITAE) by 61.5%. This work validates the MRAC as a powerful and robust solution for robots operating in unpredictable environments.**

***Keywords*—Model reference adaptive control; Mecanum Wheel Robot; center of gravity offset**

I. Introduction

Omnidirectional mobile robots equipped with Mecanum wheels are integral to modern logistics, manufacturing and service applications that require high maneuverability in constrained environments [1], [2], [3]. However, achieving precise trajectory tracking on these platforms is hindered by the dynamic uncertainties encountered in practice. Two particularly consequential factors are unknown shifts in the center of gravity (CoG) due to variable payloads and gravitational components with coupled disturbance torques that arise during operation on inclined terrain. [4], [5], [6]. When these conditions are present simultaneously, the performance and robustness of conventional fixed-gain controllers can be severely compromised.

Conventional approaches to this tracking problem often employ linear controllers, such as the Proportional-Integral-Derivative (PID) controller, due to their simplicity and effectiveness under nominal conditions [7]. However, the performance of these fixed-gain controllers is substantially degraded when the robot dynamic parameters deviate from the nominal model [8]. To address bounded uncertainties, robust control strategies like sliding mode control (SMC) have been proposed [9]. Although effective at rejecting bounded disturbances, SMC often induces high-frequency chattering in the control signal, leading to increased energy consumption and mechanical wear. Mitigating this chattering typically requires conservative gain tuning, which in turn sacrifices tracking performance.

Although more advanced research has focused on adaptive and predictive strategies, these methods often address dynamic uncertainties in isolation. For example, some adaptive methods successfully compensate for payload variations, but their effectiveness is often limited to demonstrations on flat terrain [10]. In contrast, other studies have developed predictive strategies for navigation on sloped surfaces but do so under the critical assumption of a known and fixed CoG [11], [12]. A key limitation in the existing literature is the failure to explicitly address the significant nonlinear coupling torques that arise when a robot with an off-center CoG operates on an inclined plane. This unresolved coupling motivates the present work, as a unified framework that handles both uncertainties simultaneously remains an open challenge.

This research builds on the theoretical framework established in [13], where an MRAC was proposed to ensure the closed-loop stability of a dynamic model incorporating CoG and slope effects. However, a significant gap often exists between theoretical stability and practical performance due to unmodeled real-world dynamics. To bridge this sim-to-real gap, this study presents a rigorous experimental validation of the controller on an FMWMR. Implemented on a Rosmaster X3 platform, the controller's performance under simultaneous CoG and slope uncertainties is quantitatively assessed against a carefully tuned PID benchmark using the root mean square error (RMSE) of the position, heading error, and the integral of time-weighted absolute error (ITAE) as metrics [14], [15], [16]. The implementation is augmented with practical measures to address hardware non-idealities, including deadzone compensation, command-rate limiting, and an adaptation soft start [17], [18], [19]. To ensure that all experiments are dynamically feasible and repeatable, smooth reference trajectories are generated using quintic splines to reduce jerk transients [20], [21].

This study addresses the identified sim-to-real gap through the experimental implementation and validation of the proposed MRAC on an FMWMR. The contributions are threefold:

- The experimental validation of the model framework

from [13] on a physical robotic platform. The controller's practical effectiveness is demonstrated under challenging conditions that involve simultaneous coupled dynamic uncertainties.

- The presentation of a methodology for practical implementation. This includes essential compensation techniques for real-world hardware non-idealities, such as motor stiction and actuator saturation, which are critical for bridging the sim-to-real gap.
- The quantitative validation of the adaptive controller against a well-tuned PID benchmark. The results provide definitive proof of substantial improvements in tracking accuracy and disturbance rejection, particularly when the system is challenged by the complex coupling of inertial and gravitational forces.

This study is organized as follows: The system modeling and adaptive controller design are summarized in Section II. The experimental setup and implementation details are described in Section III. The experimental procedure and evaluation metrics are defined in Section IV. The results and discussion are reported in Section V. Conclusions and directions for future work are provided in Section VI.

II. System Modeling and Adaptive Controller Design

This section summarizes the controller design based on the comprehensive dynamic model of the FMWMR proposed in [13]. The model captures the effects of an offset CoG and the gravitational forces on an inclined plane.

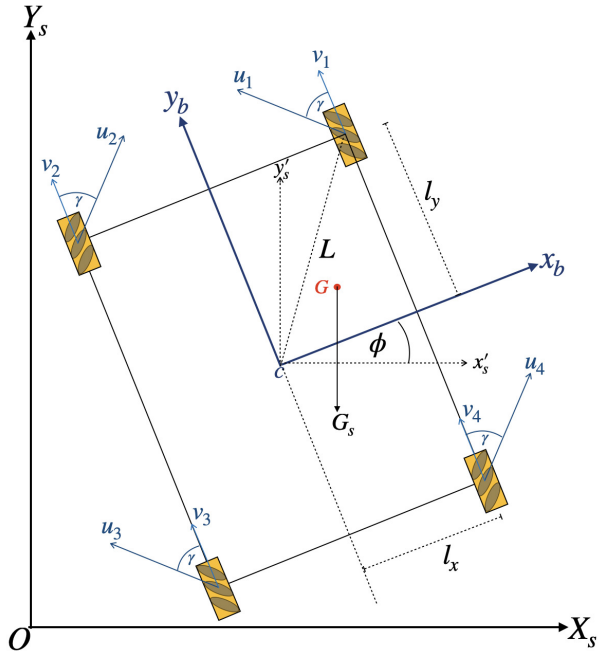


Fig. 1. A schematic of the FMWMR. The Geometric center (c), Center of gravity (G), and Wheel arrangement are indicated.

A. Kinematic Model

Consider a schematic of the FMWMR in Fig. 1. The relationship between the global frame velocity $\dot{q}$ and the body-frame velocity v_b is defined by Eq. (1):

$$\dot{q} = R(\phi)v_b, \tag{1}$$

where, $\dot{q}$ is the time derivative of the robot's pose vector in the global frame $\{X_s, O, Y_s\}$ and the rotation matrix $R(\phi)$. The robot's pose is given by the vector $q = [x, y, \phi]^{\mathrm{T}}$, where (x, y) are the Cartesian coordinates of the robot's geometric center and ϕ is the yaw angle. The body-frame velocity vector is $v_b = [v_{bx}, v_{by}, \omega_b]^{\mathrm{T}}$, which consists of the forward velocity v_{bx}, lateral velocity v_{by}, and angular velocity ω_b relative to the robot's local frame $\{x_b, c, y_b\}$. The matrix $R(\phi)$ is the rotation matrix that transforms body-frame velocities to the global frame, defined as Eq. (2):

$$R(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{2}$$

The forward kinematic model of FMWMR defines the relationship between the angular velocities of the individual wheels and the resulting linear velocities of the robot's body frame. This mapping from wheel rotational speeds ω_{wh} to body-frame velocities v_b is given by Eq. (3):

$$v_b = J\omega_{wh}, \tag{3}$$

where, $\omega_{wh} = [\omega_1, \omega_2, \omega_3, \omega_4]^{\mathrm{T}}$ is the vector of angular velocities of the four wheels and J is the Jacobian matrix given by:

$$J = \frac{R_w}{4}\begin{bmatrix} -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 \\ \frac{1}{l_x+l_y} & -\frac{1}{l_x+l_y} & -\frac{1}{l_x+l_y} & \frac{1}{l_x+l_y} \end{bmatrix},$$

R_w is the radius of the Mecanum wheels, while l_x and l_y represent the longitudinal and lateral distances from the geometric center of the robot to the center of each wheel, respectively.

B. Dynamic Model

The dynamic model of the FMWMR is derived using a Newton-Euler formulation based on two key assumptions: the robot is a rigid body, and the wheels maintain a pure rolling contact without slip. Other effects, such as actuator dynamics and viscous friction, are treated as unmodeled disturbances to be compensated for by the controller's adaptive and feedback mechanisms. The resulting dynamic model in the robot's body frame is expressed as Eq. (4):

$$H\dot{v}_b + Cv_b + G_b = F_t, \tag{4}$$

where, H is the inertia matrix that explicitly includes the CoG offset written as Eq. (5):

$$H = \begin{bmatrix} m & 0 & -my_g \\ 0 & m & mx_g \\ -my_g & mx_g & I_z \end{bmatrix}, \tag{5}$$

C is the Coriolis and centrifugal matrix and is written as Eq. (6):

$$C = \begin{bmatrix} 0 & -m\omega_b & -m\omega_b y_g \\ m\omega_b & 0 & m\omega_b x_g \\ m\omega_b y_g & -m\omega_b x_g & 0 \end{bmatrix}, \tag{6}$$

G_b is the gravitational vector and is given by Eq. (7):

$$G_b = mg\sin\alpha \begin{bmatrix} \sin\phi \\ \cos\phi \\ x_g\cos\phi - y_g\sin\phi \end{bmatrix}, \tag{7}$$

F_t is the vector of control forces and torque applied by the wheels, m is the total mass of the robot, I_z is the moment of inertia about the vertical axis passing through the CoG, g is the acceleration due to gravity, α is slope angle of the surface and (x_g, y_g) represent the coordinates of the robot's CoG within the body frame, measured relative to the geometric center.

C. Adaptive Controller Design

The control objective is to achieve asymptotic tracking of a desired trajectory, q_d, with the robot's pose, q, in the presence of parametric uncertainties within the dynamic model. The MRAC scheme is designed to meet this objective. The core of this approach is to establish a stable, reference error dynamic, which is accomplished through the formulation of a composite filtered error signal. The first step in this design is to mathematically define this error and its dynamics.

1) Error dynamics formulation: A key challenge in controlling dynamic systems is that the control inputs F_t directly influence accelerations $\dot{v}_b$ rather than the pose q. This structure creates a second-order tracking problem. To simplify this, a filtered tracking error s is introduced, which effectively transforms the second-order problem into a first-order stabilization problem. This composite signal is defined by Eq. (8):

$$s = v_b - v_r, \tag{8}$$

where, v_b is the robot's actual body-frame velocity and v_r is the reference velocity vector is strategically designed as Eq. (9):

$$v_r = R(\phi)^{\mathrm{T}}\left(\dot{q}_d - \Lambda e\right), \tag{9}$$

$e = q - q_d$ is the pose error and Λ is a positive-definite gain matrix. This construction is critical because it ensures that if the controller drives $s \to 0$, the pose error, e, is guaranteed to converge exponentially according to the stable dynamic $\dot{e} = -\Lambda e$.

The error dynamics for control design are derived by differentiating Eq. (8) and substituting the system's dynamic model, yielding Eq. (10):

$$H\dot{s} + Cs = F_t - (H\dot{v}_r + Cv_r + G_b). \tag{10}$$

This equation provides a clear structure for the control design. However, the term $(H\dot{v}_r + Cv_r + G_b)$ contains the unknown physical parameters. The next step, therefore, is to restructure this term to systematically handle these uncertainties.

2) System and control law: To address the parametric uncertainty in the system dynamics, a linear parameterization is performed. This crucial step reformulates the dynamic terms containing unknown physical properties into a structure that is linear in the parameters. This separates the dynamics into two distinct components: a vector θ containing all the unknown physical parameters and a known regressor matrix Y constructed from measurable signals of the reference trajectory. This reformulation is expressed as Eq. (11):

$$H\dot{v}_r + Cv_r + G_b = Y\theta, \tag{11}$$

where, the vector of unknown parameters θ, and known regressor Y and defined as Eq. (12):

$$\theta = [m, I_z, mx_g, my_g, m\sin\alpha, mx_g\sin\alpha, my_g\sin\alpha]^{\mathrm{T}}, \tag{12}$$

and

$$Y = \begin{bmatrix} y_{11} & y_{12} & 0 \\ 0 & 0 & \dot{\omega}_{r\phi} \\ 0 & y_{23} & y_{33} \\ y_{14} & 0 & y_{34} \\ g\sin\phi & g\cos\phi & 0 \\ 0 & 0 & g\cos\phi \\ 0 & 0 & -g\sin\phi \end{bmatrix}^{\mathrm{T}},$$

where, y_{ij} are functions of the reference trajectory and its derivatives are defined as:

$$\begin{aligned} y_{11} &= \dot{v}_{rx} - v_{ry}\omega_{r\phi}, & y_{12} &= \dot{v}_{ry} - v_{rx}\omega_{r\phi}, \\ y_{23} &= \dot{\omega}_{r\phi} + \omega_{r\phi}^2, & y_{33} &= \dot{v}_{ry} + v_{ry}\omega_{r\phi}, \\ y_{14} &= -(\dot{\omega}_{r\phi} + \omega_{r\phi}^2), & y_{34} &= -(\dot{v}_{rx} + v_{rx}\omega_{r\phi}). \end{aligned}$$

Based on this equation, the control law is formulated as Eq. (13):

$$F_t = Y\hat{\theta} - Ks, \tag{13}$$

where, $\hat{\theta}$ is the estimation of θ and K is a positive-definite feedback gain matrix. This law consists of a feed-forward term, $Y\hat{\theta}$, to cancel the estimated system dynamics, and a stabilizing feedback term -Ks. The parameter estimates are updated using the following adaptation law Eq. (14):

$$\dot{\hat{\theta}} = -\Gamma Y^{\mathrm{T}} s, \tag{14}$$

where, Γ is a positive definite matrix of adaptation gains.

With the complete control and adaptation laws now fully defined, it is necessary to formally prove that this closed-loop system is stable and that the tracking objective is indeed achieved.

3) Stability analysis: The stability of the closed-loop system is demonstrated using Lyapunov's direct method. Consider the Lyapunov candidate function, Eq. (15):

$$V = \frac{1}{2}s^{\mathrm{T}}Hs + \frac{1}{2}\tilde{\theta}^{\mathrm{T}}\Gamma^{-1}\tilde{\theta}, \tag{15}$$

where, $\tilde{\theta} = \theta - \hat{\theta}$ is the parameter estimation error. The function V is positive definite as it is a quadratic form of the tracking and estimation errors. The time derivative of V, after substituting the error dynamics and using the skew-symmetric property of the matrix $(\dot{H} - 2C)$, simplifies as Eq. (16):

$$\dot{V} = s^{\mathrm{T}}(F_t - Y\theta) + \tilde{\theta}^{\mathrm{T}}\Gamma^{-1}\dot{\tilde{\theta}}. \tag{16}$$

Substituting the control law in Eq. (13) and adaptation law in Eq. (14), and noting that $\dot{\tilde{\theta}} = -\dot{\hat{\theta}}$, yields the final expression for the derivative as Eq. (17):

$$\dot{V} = -s^{\mathrm{T}}Ks. \tag{17}$$

Since K is positive-definite, $\dot{V} \leq 0$, proving that V is nonincreasing and that s and $\tilde{\theta}$ are bounded. By Barbalat's Lemma, it can be further shown that $s \to 0$ as $t \to \infty$. As established in the error dynamics formulation, this ensures the asymptotic convergence of the pose tracking error, thus rigorously satisfying the overall control objective.

D. Control Architecture Overview

The control architecture, illustrated in Fig. 2, operates through two nested loops. First, an outer loop determines the pose error e between the desired trajectory and the robot's actual pose. This error is processed by the Reference Velocity according to Eq. (9) to produce a reference velocity designed to guide the robot. Next, an inner loop computes the filtered error signal, as defined by Eq. (8), by comparing the robot's actual body velocity to this reference velocity. The filtered error signal is central to the adaptive mechanism, driving both the Adaptation Law to update parameter estimates $\hat{\theta}$ and the control law to calculate the final control force F_t. This closed-loop structure allows the controller to drive the tracking error to zero while simultaneously learning and compensating for the unknown dynamic forces in real-time.

III. Experimental Setup

This section details the experimental framework used to validate the performance and robustness of the MRAC. The hardware platform and state estimation methods are designed to facilitate rigorous testing under significant variations in the robot's payload distribution.

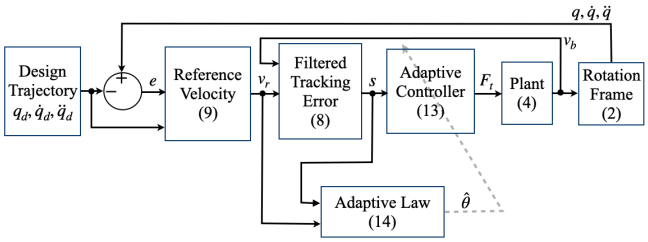


Fig. 2. Block diagram of the proposed MRAC framework.

A. Hardware Platform

The experiments are conducted on a Rosmaster X3, the FMWMR, as depicted in Fig. 3.

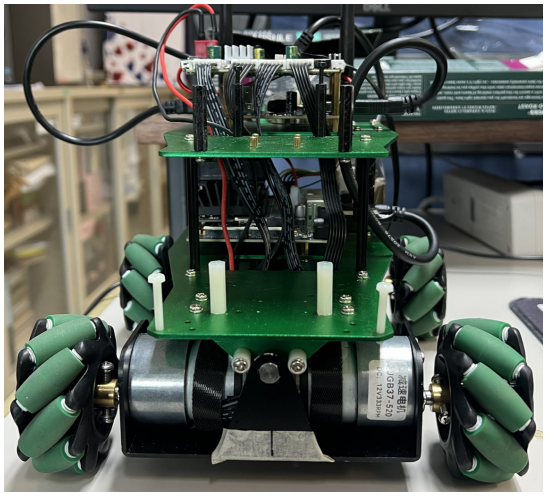


Fig. 3. The Rosmaster X3 experimental platform.

The platform's key physical parameters are summarized in Table I.

TABLE I. Rosmaster X3 Robot Parameters

Parameter	Symbol	Value
Robot Mass (Base)	m_r	2.0 [kg]
Added Mass	m_b	0.4 [kg]
Half Wheelbase	l_x	0.085 [m]
Half Track Width	l_y	0.080 [m]
Mecanum Wheel Radius	R_w	0.0325 [m]
Motor Encoder Resolution	N_{enc}	1320 [PPR]

The robot is equipped with two primary sensor systems for state estimation: four high-resolution quadrature wheel encoders fitted to the DC motors and an onboard IMU.

These sensors provide the raw data necessary for estimating the robot's pose and velocity, a process which is detailed next.

B. State Estimation and Signal Processing

Accurate estimation of the robot's state, comprising its world-frame pose and body-frame velocities, is essential for feedback control. This state is derived by processing and fusing data from the onboard sensors.

Estimation of body-frame velocities is a two-part process. First, the linear components are derived from the motor encoders. The angular velocity of each individual wheel is calculated from its respective motor encoder as in Eq. (18):

$$\omega_i = \frac{2\pi \Delta N_i}{N_{enc} \Delta t}, \tag{18}$$

where, ΔN_i is the encoder pulse count over the sampling interval Δt. These individual wheel velocities are then assembled into a vector and transformed into the robot's body-frame linear velocity v_b using the inverse kinematic model from Eq. (3). To ensure accuracy and mitigate noise, the body-frame angular velocity, ω_b, is derived by numerically differentiating the filtered yaw measurement from the IMU.

Next, the world-frame pose is estimated. The robot's orientation, ϕ, is obtained directly from the IMU's yaw measurement, with an initial offset recorded to define the world frame. The world-frame position (x, y) is then calculated by rotating the body-frame velocity estimate, v_b, into the world frame and performing first-order Euler integration at each time step. This process constitutes the robot's odometry.

Finally, to mitigate measurement noise before use in the control loop, the estimated state derivatives are processed with a first-order digital low-pass filter [19]. The filter is implemented by Eq. (19):

$$y_k = y_{k-1} + \beta(x_k - y_{k-1}), \tag{19}$$

where, y_k is the filtered output at the current time step, x_k is the raw input signal and β is the filter coefficient and calculated from the filter time constant τ as Eq. (20):

$$\beta = \frac{\Delta t}{\tau + \Delta t}. \tag{20}$$

The selection of the time constant, τ, presents a critical trade-off between noise rejection and phase lag, which can impact control stability. Based on empirical tuning, a time constant of $\tau = 0.04$ [s] is selected for the relatively clean IMU signal to prioritize a fast response. A slightly larger constant of $\tau = 0.05$ [s] is applied to the noisier, encoder-derived linear velocities to provide more aggressive filtering.

With a reliable state estimate available for feedback, the controller requires a physically achievable reference trajectory to follow. The generation of these smooth trajectories is a critical prerequisite for high-performance control.

C. Generation of Smooth Reference Trajectories

The performance of the controller is fundamentally limited by the quality of the reference trajectory. It is commanded to follow. For a physical system such as an FMWMR, commanding instantaneous changes in velocity or acceleration is impossible and leads to detrimental effects such as wheel slip, actuator saturation, and mechanical vibrations. Therefore, ensuring that the trajectory is sufficiently smooth is not an optimization, but a critical requirement to achieve high-fidelity motion.

1) Reference trajectory design: High-fidelity motion control requires reference trajectories that are physically realizable. Trajectories with discontinuous acceleration degrade tracking performance by inducing wheel slip and actuator saturation. Therefore, all reference signals are generated using quintic polynomial splines to ensure smoothness.

Unlike simpler cubic splines, quintic splines provide C^2-continuity, which guarantees a continuous acceleration profile. This is achieved by defining six boundary conditions: initial and final position (p_i, p_f), velocity (v_i, v_f), and acceleration (a_i, a_f). By setting the initial and final accelerations to zero, the jerk is eliminated, ensuring smooth rest-to-rest maneuvers as illustrated in Fig. 4. Here, Fig. 4(a) details the smooth S-curve position profile, Fig. 4(b) the corresponding bell-shaped velocity profile, and Fig. 4(c) the C^2-continuous acceleration. The acceleration begins and ends at zero, a condition which eliminates jerk and ensures the smoothness of the entire trajectory.

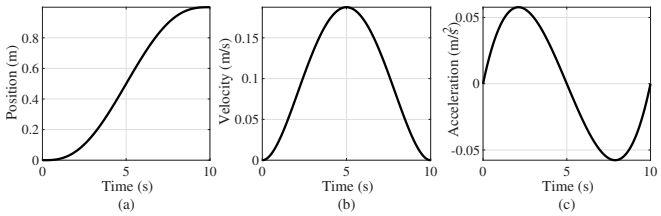


Fig. 4. Kinematic profiles of a 1 [m] maneuver using a quintic spline. (a) The smooth S-curve position profile, (b) The corresponding bell-shaped velocity profile, (c) The C^2-continuous acceleration.

The position profile $p(t)$ for any single degree of freedom is defined by the quintic polynomial as Eq. (21):

$$p(t) = c_0 + c_1 t + c_2 t^2 + c_3 t^3 + c_4 t^4 + c_5 t^5, \tag{21}$$

where, the six coefficients $c = [c_0, \ldots, c_5]^{\mathrm{T}}$ are uniquely determined by Eq. (22):

$$\begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 1 & T & T^2 & T^3 & T^4 & T^5 \\ 0 & 1 & 2T & 3T^2 & 4T^3 & 5T^4 \\ 0 & 0 & 2 & 6T & 12T^2 & 20T^3 \end{bmatrix}^{-1} \begin{bmatrix} p_i \\ v_i \\ a_i \\ p_f \\ v_f \\ a_f \end{bmatrix}, \tag{22}$$

where, T is total trajectory time.

Solving this system yields the unique set of coefficients required to generate the smooth trajectory for each experimental path. This method is used to generate separate, smooth trajectories for each experimental path. While this method produces ideal reference signals, deploying the theoretical controller on the physical hardware requires additional practical considerations to compensate for real-world non-idealities.

D. Practical Implementation and Non-Ideality Compensation

To ensure robust performance on the physical hardware, the theoretical MRAC framework is augmented with three enhancements to compensate for non-ideal dynamics: friction, command saturation, and aggressive initial adaptation.

1) Feed-forward compensation: Nonlinear friction, particularly stiction, introduces a deadzone that degrades low-velocity tracking performance. To counteract this, a continuous, model-based feed-forward compensator is implemented using the hyperbolic tangent function to avoid control chattering as Eq. (23):

$$u_{ff} = \begin{bmatrix} V_{dz,x} \tanh\left(\frac{v_{rx}}{\epsilon_v}\right) \\ V_{dz,y} \tanh\left(\frac{v_{ry}}{\epsilon_v}\right) \\ V_{dz,\omega} \tanh\left(\frac{\omega_{r\phi}}{\epsilon_\omega}\right) \end{bmatrix}, \tag{23}$$

In this formulation, the critical parameter V_{dz} represents the breakaway voltage, which corresponds to the minimum voltage required to overcome stiction. The terms $V_{dz,x}$, $V_{dz,y}$, and $V_{dz,\omega}$ are the specific breakaway voltages for the longitudinal, lateral, and rotational axes. The parameters ϵ_v and ϵ_ω are smoothing factors that dictate the sharpness of the compensation signal's transition. All parameters are empirically tuned to improve low-velocity response and minimize steady-state error without introducing oscillations.

In addition to static friction, another practical limitation is the physical saturation of motor commands, which can degrade tracking performance, particularly during aggressive maneuvers.

2) Goal-proximity-based rate limiting: Fixed acceleration limits impose an inherent trade-off between a robot's transit speed and its final positioning accuracy. High acceleration values reduce travel time but increase the risk of overshooting the target, while low values ensure precision at the cost of efficiency. To resolve this, a state-dependent rate-limiting strategy is implemented, which dynamically modulates the acceleration limit, a_{lim}, based on the robot's proximity to its target d_{goal}. The modulation is performed in a two-step process. First, a unitless scaling factor, k, is calculated based on the robot's position within a pre-defined deceleration zone, as in Eq. (24):

$$k = \text{clip}\left(\frac{d_{goal} - d_{min}}{d_{max} - d_{min}}, 0, 1\right), \tag{24}$$

where, d_{max} and d_{min} define the outer and inner boundaries of the deceleration zone. As the robot travels through this zone, the value of k transitions linearly from 1 to 0, with the clip function, $\text{clip}(X, lower, upper)$ clips the values in array X to the range $[lower, upper]$ by replacing values less than lower with lower and values greater than upper with upper, that ensures it remains bounded.

This scaling factor is then used to linearly interpolate the final acceleration limit a_{lim} between two prescribed bounds, as in Eq. (25):

$$a_{lim} = a_{min} + k \cdot (a_{base} - a_{min}), \tag{25}$$

where, a_{base} represents the high baseline acceleration for rapid transit (applied when $k \approx 1$), while a_{min} is the conservative minimum acceleration for precise final movements (applied when $k \approx 0$). This method ensures a smooth transition from aggressive to conservative motion, optimizing both speed and accuracy.

Beyond compensating for these physical non-linearities, the controller's transient behavior during startup must also be managed to ensure stability.

3) Adaptation law for soft start: A significant initial tracking error can induce large, abrupt updates to the adaptive parameters $\hat{\theta}$, potentially causing control signal saturation or transient instability. To mitigate this, a soft-start mechanism is implemented by scheduling the adaptation gain matrix Γ. The gain is linearly ramped from an initial value of zero to its nominal steady-state value according to the following rule Eq. (26):

$$\Gamma(t) = \Gamma_0 \cdot \min\left(1.0, \frac{t}{T_{ramp}}\right), \tag{26}$$

where, $\Gamma(t)$ is the time-varying adaptation gain matrix, Γ_0 is its nominal value and T_{ramp} is the total duration of the ramp-up period.

The ramp duration T_{ramp} is determined empirically, selected as the minimum time required to suppress undesirable transients in the initial control signal and ensure a smooth initialization of the adaptation process.

For reproducibility, the empirically determined values for these implementation parameters are provided in Table II.

TABLE II. EMPIRICALLY TUNED IMPLEMENTATION PARAMETERS

Category	Parameter	Value
Friction Compensation		
	Deadzone Voltage $V_{dz,x}, V_{dz,y}, V_{dz,\omega}$	0.1 V
	Smoothing Factor $\epsilon_v, \epsilon_\omega$	0.1
Command Rate Limiting		
	Base Acceleration a_{base}	2.0 m/s^2
	Minimum Acceleration a_{min}	0.5 m/s^2
	Max Deceleration Distance d_{max}	0.4 m
	Min Deceleration Distance d_{min}	0.1 m
Adaptation Soft Start		
	Adaptation Ramp Time T_{ramp}	0.1 s

IV. EXPERIMENTAL PROCEDURE AND EVALUATION

A. Experimental Scenarios

To rigorously evaluate the controller's performance and robustness, experiments are conducted on a custom-built testbed, as shown in Fig. 5. The experimental design systematically varies three key factors: path type, surface condition, and payload distribution. These factors and their corresponding conditions are summarized in Table III, and the path waypoints are detailed in Table IV.

To ensure statistical reliability, each of the twelve resulting experimental scenarios is repeated three times for both the proposed MRAC and a baseline PID controller. The performance metrics presented in the results are the arithmetic mean of these three trials.

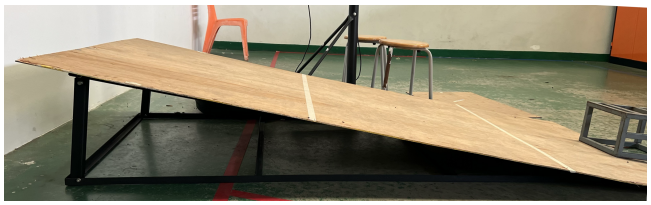


Fig. 5. The experimental testbed (2.4 m × 1.2 m), shown in its inclined configuration.

B. Center of Gravity Manipulation

To isolate the dynamic effects of an off-center payload, two CoG conditions are established without altering the robot's total mass.

- Centered CoG (Nominal): The stock robot's CoG is not at its geometric center. A baseline condition is established by adding a balancing mass, with its position empirically adjusted to ensure equal weight distribution across all four wheels.
- Off-center CoG: For this condition, the same balancing mass is relocated to a fixed position to induce a significant, known CoG shift, as specified in Table III.

This methodology ensures that any observed performance differences between the two payload conditions are attributable solely to the change in dynamic coupling caused by the CoG shift, not by a change in total mass.

TABLE III. EXPERIMENTAL FACTORS AND CONDITIONS

Factor	Conditions
Path Type	Straight, Diagonal, Square
Surface Condition	1. Flat Surface 2. Inclined Surface ($\pi/18$ rad)
Payload Distribution	1. Centered CoG (Nominal) 2. Off-center CoG $(x_g, y_g) = (-4.5, -26.7)$ mm

TABLE IV. REFERENCE TRAJECTORY DEFINITIONS

Path Name	Description	Waypoints (x, y) [m]
Straight	1 m translation along y-axis	(0,0) → (0,1)
Diagonal	1 m translation in x and y	(0,0) → (1,1)
Square	4 m perimeter path with sharp corners	(0,0) → (0,1) → (1,1) → (1,0) → (0,0)

C. Benchmark Controller for Comparison

For performance comparison, a fixed-gain PID controller is implemented as a benchmark. The controller's gains are carefully tuned on the physical hardware. Initial estimates are obtained using the Ziegler-Nichols method and subsequently optimized using MATLAB's PID Tuner with experimental step response data. The final, optimized gains for the translational and rotational axes are presented in Table V.

TABLE V. OPTIMIZED PID CONTROLLER GAINS

Control Axis	k_p	k_i	k_d
Translational (x, y)	2.80	0.06	0.15
Rotational (ϕ)	2.20	0.04	0.12

D. Controller Gain Tuning Methodology

The performance of MRAC is determined by three gain matrices of the error convergence matrix Λ, the feedback matrix K and the adaptation matrix Γ. A systematic two-phase empirical tuning procedure is conducted on the physical hardware to achieve a balance between responsiveness and stability. The final gains are listed in Table VI.

1) Phase 1: Tuning baseline feedback gains Λ and K: First, a stable, non-adaptive baseline controller is established by setting the adaptation gain $\Gamma = 0$ and initializing the parameter vector $\hat{\theta}$ with nominal estimates. The gains in $\Lambda = \text{diag}(\lambda_p, \lambda_p, \lambda_\phi)$ are tuned to set the desired error convergence rate without causing oscillation. Subsequently, the feedback gains $K = \text{diag}(k_p, k_p, k_\phi))$ are tuned to ensure a critically damped response without amplifying sensor noise or causing jerky commands.

2) Phase 2: Tuning adaptation gains Γ: Second, the adaptation gains in the diagonal matrix $\Gamma = \text{diag}(\gamma_1, \ldots, \gamma_7)$ are tuned to set the parameter learning rates. The selection of these gains presents a critical trade-off: Low values result in slow adaptation, while high values can cause noisy parameter estimates and instability.

The gains are tuned incrementally, starting with small values. They are increased to the highest level that provided rapid error convergence without introducing oscillations into the control signal. This process is performed sequentially using maneuvers designed to excite specific dynamics, for example, tuning mass/inertia gains during linear accelerations and CoG/slope gains during tests with an off-center payload on the inclined surface.

TABLE VI. MRAC GAIN PARAMETERS

Gain Parameter	Value
Error Convergence Gains (Λ)	
Translational Gain λ_p	2.5
Rotational Gain λ_ϕ	2.5
Control Gain (K)	
Translational Gain k_p	0.5
Rotational Gain k_ϕ	0.5
Adaptation Gains (Γ)	
γ_1 for m	0.0400
γ_2 for I_z	0.0180
γ_3 for mx_g	0.0120
γ_4 for my_g	0.0010
γ_5 for $m \sin\alpha$	0.0050
γ_6 for $mx_g \sin\alpha$	0.0006
γ_7 for $my_g \sin\alpha$	0.0006

E. Performance Metrics

To provide a quantitative assessment of controller performance, two standard metrics are chosen. One is the RMSE to evaluate tracking accuracy and the other is the ITAE to evaluate

settling performance. The specific equations are defined as Eq. (27), Eq. (28) and Eq. (29):

$$\text{RMSE}_p = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_p(t_i)^2}, \tag{27}$$

$$\text{RMSE}_h = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_h(t_i)^2} \tag{28}$$

and

$$\text{ITAE}_p = \int_0^T t \cdot |e_p(t)|\, dt, \tag{29}$$

where, $e_p(t) = p(t) - p_d(t)$ is the error between actual position $p(t)$, desired position $p_d(t)$ and $e_h(t) = \phi(t) - \phi_d(t)$ and N is total number of data points.

V. Results and Discussion

A. Experimental Results

This section presents the experimental results that compare the proposed MRAC with the baseline PID controller in all 12 test scenarios. To highlight the controllers' distinct behaviors, a detailed graphical analysis is presented for two representative scenarios: the nominal case (straight-line motion, flat surface, centered payload) and the most challenging case (square-path motion, inclined surface, off-center payload). A comprehensive summary of the performance metrics for all experiments is provided at the end of this section.

1) Performance in nominal conditions: The experimental results for a straight line trajectory 1 [m] under nominal conditions are shown in Fig. 6. Performance comparison between MRAC and PID controllers is shown in Fig. 6(a) to Fig. 6(d), where the ideal trajectory is represented by a solid black line, the PID response by a blue dashed line, and the MRAC response by a red dotted line. Fig. 6(a) shows the robot's lateral position, where both responses remain close to the ideal zero position. Fig. 6(b) shows the primary translation, which follows the smooth S-shaped reference trajectory. The robot's orientation is shown in Fig. 6(c), where the desired zero-radian heading is successfully maintained. A top-down view of the path is provided in Fig. 6(d). Fig. 6(e) shows the RMSE of position over time. The final RMSE for the MRAC is approximately 1.5 [cm], a 25% reduction compared to the PID's 2.0 [cm] error. Finally, Fig. 6(f) confirms that the heading error for both controllers remains minimal and stable throughout the maneuver.

2) Performance under challenging conditions: The experimental results for a square path trajectory under challenging conditions (an inclined surface with an off-center CoG) are presented in Fig. 7. The robot's position along the x- and y-axes is shown in Fig. 7(a) and Fig. 7(b), where the MRAC's red dotted line more closely follows the sharp transitions of the reference compared to the PID's response. The robot's orientation is shown in Fig. 7(c). A top-down view of the path is provided in Fig. 7(d), which illustrates that the PID controller exhibits significant corner overshoot, while the MRAC maintains sharp, accurate tracking. The RMSE of position in Fig. 7(e) highlights the performance difference; the PID's error accumulates, reaching a final RMSE of approximately 2.5 [cm], whereas the MRAC maintains a stable and bounded error of just 1.0 [cm]. Finally, Fig. 7(f) confirms that the MRAC's heading error profile is visibly smoother than the PID's.

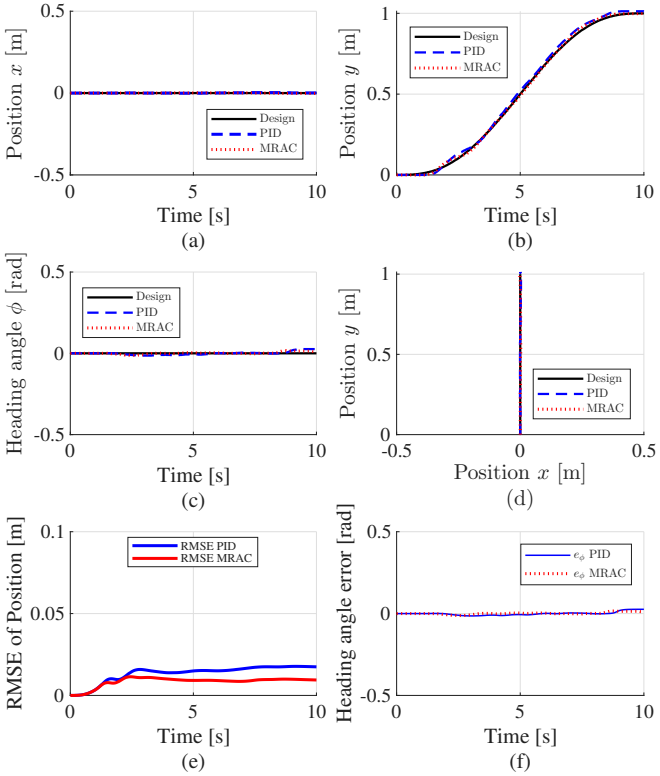


Fig. 6. Straight path tracking results for the nominal case comparing PID and MRAC controllers: (a) Position x over time. (b) Position y over time. (c) Heading angle ϕ over time. (d) Robot's trajectory in the $x-y$ plane. (e) RMSE of position over time. (f) Heading angle error over time.

3) Quantitative performance summary: The complete quantitative results from all twelve experimental scenarios are compiled in Table VII. For each path type, surface condition, and CoG configuration, the table lists the position RMSE, heading RMSE, and position ITAE for both the PID and MRAC controllers. The best-performing result in each case is highlighted in bold.

B. Discussions

The experimental results presented in Section V-A consistently demonstrate the superior performance and robustness of the MRAC framework compared to a well-tuned PID controller. The quantitative data in Table VII reveals that, on average, the MRAC reduced the position RMSE by 52.7% and the position ITAE by 61.5%. This substantial improvement can be directly attributed to the controller's adaptive nature.

The key strength of the MRAC lies in its ability to estimate and compensate for dynamic uncertainties in real time, a capability that fixed-gain controllers lack. This is particularly evident in the most challenging scenarios. For instance, in the

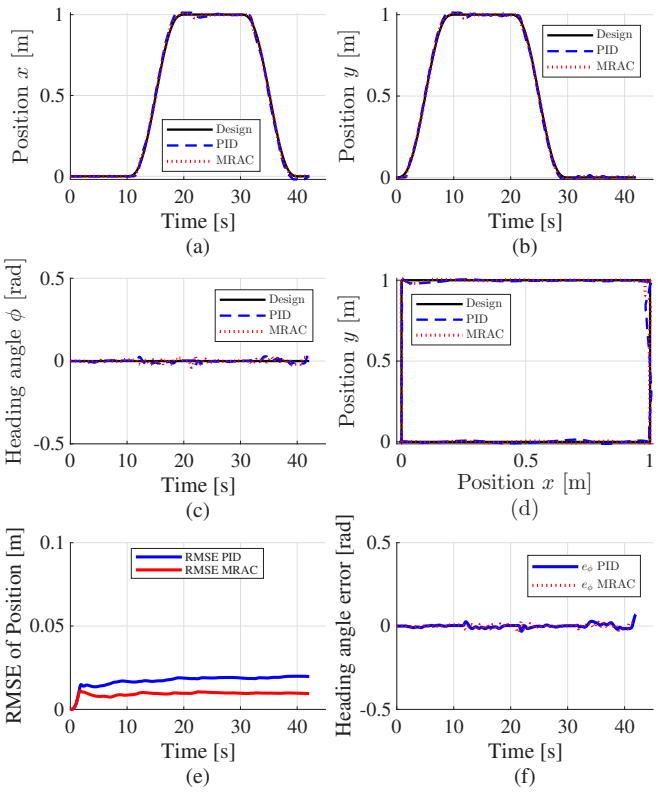


Fig. 7. Square path tracking results under challenging conditions comparing PID and MRAC controllers: (a) Position x over time. (b) Position y over time. (c) Heading angle ϕ over time. (d) Robot's trajectory in the $x-y$ plane. (e) RMSE of position over time. (f) Heading angle error over time.

square path test with an off-center CoG on an inclined surface, the PID controller's performance degraded significantly, as shown by the large corner overshoots in Fig. 7(d). This occurs because fixed PID gains cannot counteract complex coupled disturbance torques that arise from the combination of CoG offset and gravitational forces. In contrast, the MRAC uses the regressor matrix Y and the adaptation law in Eq. (14) to continuously update its internal model, which effectively cancels these disturbances and maintains high-precision tracking.

Furthermore, the MRAC's robustness to internal dynamic coupling was highlighted during the diagonal path tests. The PID controller struggled to manage the simultaneous translational and rotational motion, resulting in a heading RMSE of 0.0947 [rad]. In contrast, the MRAC achieved a heading RMSE of just 0.0131 [rad], an 86% reduction. This result demonstrates the effectiveness of the adaptation mechanism in decoupling the system dynamics, which is a critical feature for holonomic platforms like the FMWMR.

The controller's results are highly repeatable, as confirmed by running each experiment three times. The statistical variability is very low, with the standard deviation of the error typically remaining below 5% of the mean value. A sensitivity analysis further confirmed the controller's robustness to parameter tuning. Moderate changes (approximately $\pm 20\%$) to the controller gains resulted in graceful performance degradation rather than instability, proving that its effectiveness is an inherent feature of the adaptive design, not the result of fragile,

TABLE VII. COMPARATIVE PERFORMANCE METRICS FOR ALL PATH TYPES

Path	Condition Surface/CoG	Controller	RMSE_p [m]	RMSE_h [rad]	ITAE_p [m s]
Straight	Flat/Center	PID	0.0170	0.0189	23.9470
		MRAC	**0.0094**	**0.0105**	**9.4366**
	Flat/Offset	PID	0.0191	0.0299	21.0450
		MRAC	**0.0100**	**0.0259**	**9.7486**
	Slope/Center	PID	0.0164	0.0369	20.0144
		MRAC	**0.0061**	**0.0367**	**5.8353**
	Slope/Offset	PID	0.0163	0.0346	20.4149
		MRAC	**0.0091**	**0.0305**	**11.9156**
Diagonal	Flat/Center	PID	0.0236	0.0947	37.6089
		MRAC	**0.0106**	**0.0131**	**12.0401**
	Flat/Offset	PID	0.0232	0.0573	37.4503
		MRAC	**0.0111**	**0.0477**	**13.2413**
	Slope/Center	PID	0.0234	0.0949	35.7405
		MRAC	**0.0075**	**0.0348**	**7.8987**
	Slope/Offset	PID	0.0222	0.0863	37.7239
		MRAC	**0.0117**	**0.0719**	**13.4543**
Square	Flat/Center	PID	0.0191	0.0387	396.1914
		MRAC	**0.0092**	**0.0165**	**158.7575**
	Flat/Offset	PID	0.0192	0.0391	406.1247
		MRAC	**0.0085**	**0.0168**	**148.7291**
	Slope/Center	PID	0.0191	0.0300	396.3069
		MRAC	**0.0093**	**0.0268**	**187.6281**
	Slope/Offset	PID	0.0197	0.0182	417.5931
		MRAC	**0.0095**	**0.0150**	**167.6022**

hyperspecific tuning.

To address the practical trade-offs between the two controllers, a comparison of their computational requirements is provided in Table VIII. The analysis highlights that the PID controller's logic is based on simple algebraic operations, while the MRAC is inherently more complex, requiring the assembly of a regressor matrix and several matrix-vector multiplications in each cycle. We measured the average execution time for a single control loop on the Rosmaster X3's onboard processor. The results show that the PID controller took approximately 0.08 [ms] per cycle, while the more demanding MRAC took 0.27 [ms]. Although this makes the MRAC about 3.4 times more computationally intensive, its execution time is still extremely low, allowing for a theoretical operating frequency of over 3,700 Hz. This rate is far higher than the mechanical limits of the robot, confirming that the proposed MRAC is a highly efficient and practical solution for real-time control.

In summary, the experimental validation confirms that the proposed MRAC is not only stable in theory but also highly effective in practice. By explicitly addressing the sim-to-real gap with practical compensations and leveraging a sound adaptive control law, the controller provides a robust solution for high-precision motion in unpredictable environments.

VI. CONCLUSION

An adaptive controller for an FMWMR with unknown CoG offset and unknown slope inclination has been implemented and experimentally validated on a Rosmaster X3 platform. A linearly parameterized dynamic model that captures CoG-induced coupling and gravity on an incline is employed, and

TABLE VIII. COMPUTATIONAL COMPARISON OF PID AND MRAC

Aspect	PID Controller	MRAC
Core Calculation	- Error calculation - Summation (Integral) - Differentiation	- Regressor matrix assembly - Matrix-vector multiplication
Parameter Updates	Fixed offline-tuned gains	Online adaptation law via gradient descent
Memory Usage	Low (stores error states)	Moderate (stores error states and parameter vector $\hat{\theta}$)
Avg. Execution Time*	0.08 [ms]	0.27 [ms]

*Average per-cycle execution time measured on the Rosmaster X3.

stability is ensured through a Lyapunov-based adaptation law. To bridge the gap between theory and experiment, deployment-oriented measures that address motor deadzones, command rate limits, and adaptation transients are incorporated. Across 12 scenarios, consistent improvements are demonstrated over a carefully tuned PID benchmark. On average, the position RMSE is reduced by 52.7% and the position ITAE is reduced by 61.5%. Heading accuracy is kept bounded and precise even under the most challenging combinations of slope and CoG shift. These results indicate that the controller achieves high-precision tracking under coupled uncertainties in real time, demonstrating its suitability for operation in unstructured environments.

Future work will focus on improving the robot's long-term autonomy and robustness by addressing three key areas. The current reliance on internal sensors, while effective for short maneuvers, is susceptible to cumulative drift. To overcome this, our primary goal is to develop a robust global localization system by integrating a 2D LiDAR sensor and implementing a Simultaneous Localization and Mapping (SLAM) algorithm, with an Extended Kalman Filter (EKF) for sensor fusion. Building on this improved state estimation foundation, we will then rigorously test and adapt the controller for more complex and unstructured terrains, including surfaces with varying friction and irregularities, to potentially extend the adaptive model to account for wheel slip. Ultimately, the enhanced controller and localization system will be integrated as the low-level execution layer within a complete navigation stack. By coupling our work with a high-level motion planner capable of real-time environmental perception, we aim to transition from trajectory tracking to full autonomy, enabling the robot to independently generate and modify paths to safely navigate around obstacles in real-world applications.

ACKNOWLEDGMENT

This work is partially supported by JSPS KAKENHI Grant Number JP21K03930 and (C) 23K03898.

REFERENCES

[1] R. Siegwart, I. R. Nourbakhsh, and D. Scaramuzza, *Introduction to Autonomous Mobile Robots*, 2nd ed. Cambridge, MA, USA: MIT Press, 2011.

[2] G. Klančar, A. Zdešar, S. Blažič, and I. Škrjanc, *Wheeled Mobile Robotics: From Fundamentals Towards Autonomous Systems*. Oxford, U.K.: Butterworth-Heinemann, 2017.

[3] T.-N. Manh, M.-P. Xuan, P.-N. Doan, and T.-P. Quoc, "Tracking control for mobile robots with uncertain parameters based on model reference adaptive control," in *Proc. Int. Conf. Control, Autom. Inf. Sci. (ICCAIS)*, Danang, Vietnam, Nov. 2013, pp. 18–23.

[4] Z. Hendzel and Ł. Rykała, "Modelling of dynamics of a wheeled mobile robot with Mecanum wheels with the use of Lagrange equations of the second kind," *Int. J. Appl. Mech. Eng.*, vol. 22, no. 1, pp. 81–99, Mar. 2017.

[5] I. Zeidis and K. Zimmermann, "Dynamics of a four-wheeled mobile robot with Mecanum wheels," *Z. Angew. Math. Mech.*, vol. 99, no. 12, Art. no. e201900173, Dec. 2019.

[6] S. O. Onyango, Y. Hamam, K. Djouani, and G. Qi, "Dynamic control of powered wheelchair with slip on an incline," in *Proc. 2nd Int. Conf. Adapt. Sci. Technol. (ICAST)*, Accra, Ghana, Jan. 2009, pp. 278–283.

[7] N. H. Thai, T. Ly, and L. Dzung, "Trajectory tracking control for mecanum wheel mobile robot by time-varying parameter PID controller," *Bull. Electr. Eng. Inform.*, vol. 11, no. 4, pp. 1902–1910, Aug. 2022.

[8] J.-J. E. Slotine and W. Li, *Applied Nonlinear Control*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1991.

[9] X. Lu, X. Zhang, G. Zhang, and S. Jia, "Design of adaptive sliding mode controller for four-Mecanum wheel mobile robot," in *Proc. 37th Chinese Control Conf. (CCC)*, Wuhan, China, Jul. 2018, pp. 3983–3987.

[10] V. Alakshendra and S. Chiddarwar, "Adaptive robust control of Mecanum-wheeled mobile robot with uncertainties," *Nonlinear Dyn.*, vol. 87, no. 4, pp. 2147–2169, Feb. 2017.

[11] X. Yue, J. Chen, Y. Li, *et al.*, "Path tracking control of skid-steered mobile robot on the slope based on fuzzy system and model predictive control," *Int. J. Control Autom. Syst.*, vol. 20, no. 4, pp. 1365–1376, Apr. 2022.

[12] H. B. Santos *et al.*, "Model predictive torque control for velocity tracking of a four-wheeled climbing robot," *Sensors*, vol. 20, no. 24, Art. no. 7059, Dec. 2020.

[13] C. Chaichumporn *et al.*, "The dynamical modeling of four mecanum wheel mobile robot on typical unstructured terrain," in *Proc. 21st Int. Conf. Electr. Eng./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON)*, Khon Kaen, Thailand, May 2024, pp. 1–5.

[14] K. Ogata, *Modern Control Engineering*, 5th ed. Upper Saddle River, NJ, USA: Prentice Hall, 2010.

[15] Y. Luo *et al.*, "Trajectory tracking control of an orchard robot based on improved integral sliding mode algorithm," *Agriculture*, vol. 15, no. 17, p. 1881, 2025.

[16] A. D. Abdullsahiba and H. M. Alwan, "Experimental investigation of applying multi-stage (PID-MRAC) controller for trajectory tracking of four mecanum wheeled mobile robot," *Eng. Technol. J.*, vol. 43, no. 08, pp. 641–658, 2025.

[17] S. Ibrayev and B. Omarov, "Designing an autonomous mobile robot featuring self-localization through 3D LiDAR technology," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 5, pp. 28224–28231, Oct. 2025.

[18] M. Burkacki, I. Łysy, S. Suchoń, M. Chrzan, and R. Kowolik, "Systematic review of Mecanum and Omni wheel technologies for motor impairments," *Appl. Sci.*, vol. 15, no. 9, Art. no. 4773, Sep. 2025.

[19] G. F. Franklin, J. D. Powell, and M. Workman, *Digital Control of Dynamic Systems*, 3rd ed. Menlo Park, CA, USA: Addison-Wesley, 1998.

[20] L. Biagiotti and C. Melchiorri, *Trajectory Planning for Automatic Machines and Robots*. Berlin, Heidelberg: Springer, 2008.

[21] W. A. Farag and M. M. Mahmoud, "Trajectory optimization for autonomous highway driving using quintic splines," *World Electr. Veh. J.*, vol. 16, no. 8, p. 434, 2025.

Enhancing Out-of-Distribution Detection for Retail Time-Series Data Using Entropic Methods

Nga Nguyen Thi[1], Tuan Vu Minh[2], Khanh Nguyen-Trong[3]
Intelligent Computing for Sustainable Development Laboratory (IC4SD),
Posts and Telecommunications Institute of Technology, Vietnam[1,3]
Faculty of Information Technology, Hanoi University, Hanoi, 10000, Vietnam[1,2]
Faculty of Information Technology, Posts and Telecommunications Institute of Technology,
km10 Nguyen Trai, Hanoi, 10000, Vietnam[3]

***Abstract*—Machine learning models are typically developed under the "closed-world" assumption, where training and testing data originate from a consistent distribution. However, in real-world scenarios, especially in the retail domain, this assumption can become problematic due to the frequent introduction of new products, seasonal promotions, and irregular sales events. When models encounter out-of-distribution data inputs, predictions can become overly confident or entirely incorrect. While existing out-of-distribution detection methods primarily focus on image-based datasets, challenges associated with numerical, high-dimensional, and heterogeneous retail time-series data remain largely unexplored. To address this gap, this study proposes an enhanced Entropic Out-of-Distribution Detection framework tailored specifically for dynamic retail environments. By transforming time-series sales data into spectrogram representations and leveraging the IsoMax+ loss function, our approach improves uncertainty calibration and robustness without requiring labeled out-of-distribution data or additional post-hoc calibration techniques. Experimental results, conducted on a large-scale retail dataset from Vietnam, demonstrate that the proposed Entropic Out-of-distribution detection framework significantly outperforms traditional out-of-distribution detection methods in terms of detection accuracy and inference efficiency, providing a scalable and practical solution for real-time retail applications. Our approach achieves strong performance with an F1-score of 88% and an AUC of 91%, highlighting its promising applicability across diverse business scenarios.**

***Keywords*—Out-of-Distribution Detection, entropic learning, IsoMax+ loss, time-series classification, retail forecasting, deep learning, spectrogram transformation**

I. Introduction

In conventional supervised learning settings, models are typically developed under the "closed-world" assumption, where classifiers are trained on a predefined set of known classes and assume future data will remain consistent with the training distribution. However, this assumption rarely holds in real-world scenarios, where data distributions evolve continuously and models encounter previously unseen samples. As a result, predictive accuracy deteriorates and uncertainty estimates become unreliable when facing distributional shifts [1], [2]. To maintain model reliability, robust mechanisms for detecting anomalies and unseen data—commonly known as Out-of-Distribution detection are essential. During a preliminary survey of recent OOD research, we observed that most existing approaches are developed and benchmarked primarily on image datasets. This observation raises a natural question: Can OOD detection techniques be effectively adapted to numerical, business-oriented time-series data, such as those found in retail analytics?

In retail analytics, OOD data frequently emerge from new product launches, regional demand changes, or promotional campaigns. Such shifts can lead to unstable sales patterns, causing models to generate inaccurate forecasts and poor business decisions [3]. Conventional OOD approaches, including Maximum Softmax Probability (MSP) [4] and Mahalanobis Distance [5], are computationally simple but struggle with subtle distributional drifts. Recent entropy-based methods [6], [7] have shown promise in improving calibration by quantifying prediction uncertainty, yet they often rely on post-hoc temperature scaling or external OOD datasets [8], limiting real-world scalability.

To overcome these challenges, researchers have explored information-theoretic and distance-aware frameworks. Sugiyama et al. [9] introduced density ratio estimation to capture distributional shifts, and Macêdo et al. [10] proposed the Entropic Out-of-Distribution Detection (EOOD) framework that replaces softmax scoring with a distance-regularized entropy loss.

More recently, Liang et al. [11] and Han et al. [12] further demonstrated that entropy regularization and contrastive predictive entropy can enhance OOD robustness in deep networks and time-series tasks.

These findings confirm the potential of entropy-based learning to improve detection accuracy in dynamic, high-stakes domains such as healthcare and retail [13].

1) Motivation and objective: Despite these advances, applying OOD detection to retail time-series data remains underexplored. Retail datasets are highly non-stationary, multimodal, and influenced by seasonal, geographic, and economic factors.

This study aims to develop a scalable, entropy-driven OOD framework that can adapt to evolving retail conditions without external calibration or labeled OOD data.

Such capability can enhance model reliability, reduce forecast errors, and improve business decision-making in dynamic commercial environments.

2) Main contributions: The contributions of this work can be summarized as follows:

- We propose an enhanced Entropic Out-of-Distribution Detection (EOOD) framework tailored to retail time-series data, combining entropy and distance-based uncertainty modeling for improved calibration.
- A spectrogram-based transformation is introduced to convert sales sequences into time–frequency representations, enabling convolutional architectures to capture temporal dynamics effectively.
- A labeling and preprocessing pipeline is designed to generate realistic in-distribution and out-of-distribution samples without requiring external OOD datasets or post-hoc calibration.
- Extensive experiments on large-scale Vietnamese retail data demonstrate that the proposed method outperforms conventional baselines in detection accuracy, calibration robustness, and inference efficiency.

The remainder of this study is organized as follows: Section II reviews representative OOD detection methods and their limitations. Section III details the proposed EOOD methodology, including data processing and model architecture. Section IV presents experimental results and analysis, while Section V concludes the study and outlines future research directions.

II. Related Work

Out-of-Distribution (OOD) detection has become an essential problem in reliable deep learning, aiming to identify samples that do not conform to the training data distribution [4]. Early approaches relied on confidence-based measures, such as Maximum Softmax Probability (MSP) [4] and temperature scaling or input perturbation methods as in ODIN [14]. Later, Mahalanobis distance-based techniques [3] improved robustness by modeling feature-space distributions, while IsoMax and IsoMax+ losses [10] introduced distance-aware learning for better uncertainty calibration.

Recent research has continued to refine these directions. Liang et al. [11] proposed entropy regularization to stabilize OOD decision boundaries. Han et al. [12] extended entropy-based detection to time-series data through contrastive predictive modeling, and Liu et al. [15] leveraged spectral representation learning to enhance feature robustness. Macedo et al. [16] revisited IsoMax+ for improved distance-awareness, while Yang et al. [17] provided a comprehensive survey categorizing OOD methods by analytical principles and highlighting emerging challenges.

To synthesize these developments, Table I presents a comparative summary of representative OOD detection approaches, categorizing them by type of analysis, advantages, limitations, and their relationship to the proposed EOOD framework. This overview clarifies how our method differs by emphasizing entropy–distance synergy, calibration-free detection, and adaptability to dynamic retail time-series data.

OOD detection has emerged as a critical subfield in machine learning, particularly in safety-critical and dynamic environments. Early works often relied on confidence-based metrics such as Maximum Softmax Probability (MSP) [4], which takes the highest softmax output as a confidence score. While MSP is computationally efficient, it is prone to overconfidence on unfamiliar inputs. Distance-based methods, such as the Mahalanobis Distance approach [3], measure the feature-space distance between a sample and class centroids under Gaussian assumptions. Although effective in some cases, these methods are sensitive to covariance estimation errors and tend to degrade in high-dimensional, non-isotropic settings.

To mitigate softmax overconfidence, Liang et al. [14] proposed ODIN, which combines temperature scaling and input perturbations to improve separation between in-distribution (ID) and OOD samples. ODIN improves detection accuracy but requires careful hyperparameter tuning and access to OOD-like validation data, limiting its scalability. Entropy-based scoring methods [6] address this issue by using the entropy of the predictive distribution as an uncertainty measure—low entropy indicating confident predictions for ID data, and high entropy reflecting unfamiliarity for OOD inputs. However, many still depend on post-hoc calibration [7] or labeled OOD datasets [8], which may be impractical in real-time settings.

Entropic Out-of-Distribution Detection was introduced by Macêdo et al. [10] to overcome these limitations. EOOD replaces the softmax layer with IsoMax+, a distance-aware probability formulation that computes class likelihoods based on the squared Euclidean distance between an input's feature vector and learned class centroids. This design naturally produces higher predictive entropy for OOD samples—improving separation without post-hoc adjustments or external OOD data. IsoMax+ encourages isotropic feature distributions, making the model more robust in high-dimensional and noisy data environments. Recent studies such as Jin et al. [18] further confirm the effectiveness of entropy-based and distance-aware approaches in dynamic, high-dimensional settings.

Despite its success on image benchmarks such as CIFAR10, CIFAR100, and SVHN, most prior EOOD research has been limited to static spatial structures. In contrast, retail time-series data presents unique challenges: high dimensionality, seasonal patterns, abrupt demand spikes, and sparsity across product–location pairs. Traditional OOD methods designed for images may fail to capture these temporal–seasonal anomalies effectively.

To summarize, Table I provides a synthesis of representative OOD detection methods, categorized by their analytical principles, advantages, and limitations. This overview highlights how our proposed EOOD framework improves uncertainty calibration and adaptability to dynamic retail data.

This comparative summary facilitates a clearer understanding of how the proposed EOOD method distinguishes itself from existing approaches in terms of scalability, uncertainty calibration, and adaptability to dynamic time-series data.

This study addresses this shortage by extending EOOD to retail time-series forecasting. We introduce a spectrogram-based transformation that converts 12-month revenue sequences into image-like representations, enabling the use of deep CNN architectures such as EfficientNet-B3. By integrating IsoMax+ loss, our approach preserves EOOD's calibration-free property while enhancing its ability to detect OOD patterns in dynamic, high-variance retail environments—supporting real-time anomaly detection where timely insights are crucial for operational decision-making.:

TABLE I. COMPARATIVE SUMMARY OF REPRESENTATIVE OUT-OF-DISTRIBUTION (OOD) DETECTION METHODS CATEGORIZED BY ANALYTICAL APPROACH, HIGHLIGHTING THEIR RESPECTIVE ADVANTAGES, LIMITATIONS, AND DISTINCTIONS FROM THE PROPOSED EOOD FRAMEWORK

Method	Type of Analysis	Advantages	Limitations	Comparison with the Proposed Framework
Maximum Softmax Probability (MSP) [4]	Confidence-based	Simple, fast, easy to implement	Overconfident on unseen data	Lacks uncertainty calibration
ODIN [14]	Temperature scaling and input perturbation	Improves OOD separation	Requires tuning and validation data	Higher complexity; not scalable
Mahalanobis Distance [3]	Distance-based	Captures feature-space structure	Sensitive to covariance estimation	Less effective on high-dimensional time-series
Entropic OOD Detection (EOOD) [10]	**Entropy + Distance**	**Calibration-free, robust**	**Limited to image data**	**This study extends it to retail time-series**

$$p_i = \frac{e^{-d(f(x),c_i)}}{\sum_{j=1}^{K} e^{-d(f(x),c_j)}} \tag{1}$$

where, $f(x)$ represents the feature vector of input sample x, c_i is the centroid of class i, and $d(\cdot)$ is a distance metric such as squared Euclidean distance. This formulation encourages the model to produce higher entropy for OOD samples, as they lie farher away from the class centroids. Therefore, the model shows ability to detect OOD instances without the need for additional calibration or external datasets.

Out-of-Distribution detection has emerged as a fundamental subfield in machine learning, particularly in applications where model required reliability and safety. Several approaches have been proposed to distinguish between ID and OOD inputs, with varying assumptions and mechanisms for uncertainty estimation.

One of the earliest and most often used methods is the MSP, introduced by Hendrycks and Gimpel [4]. This method uses confidence score from the softmax output as an indicator of distributional fit. While computationally efficient, MSP likely overconfident predictions on unfamiliar inputs. In case of dealing with subtle anomalies or class overlapping distributions.

To solve these limitations, ODIN was proposed by Liang et al. [14], combining temperature scaling and input perturbations to amplify the distinction between ID and OOD samples. Despite its improved detection performance, ODIN requires careful hyperparameter tuning and access to validation data, which may hinder its scalability across domains.

Another mature method is the Mahalanobis Distance based approach introduced by Lee et al. [3], which computes the distance between a test sample and class conditional distributions in the feature space. Although this method works effective in enhancing internal feature representations,it assumes Gaussianity and sensitive to covariance estimation errors particularly in high-dimensional or non-isotropic feature spaces.

More recently, entropy-based methods have gained attention for their principled handling of predictive uncertainty. In particular, Macêdo et al. [10] proposed the Entropic Out-of-Distribution Detection framework, which replaces the conventional softmax layer with IsoMax+ loss. Traditional softmax layers calculate class probabilities based on unnormalized logits, which often lead to poorly calibrated confidence estimates, as shown by Guo et al. [7]. Therefore, this method is unsuitable with uncertain estimation.

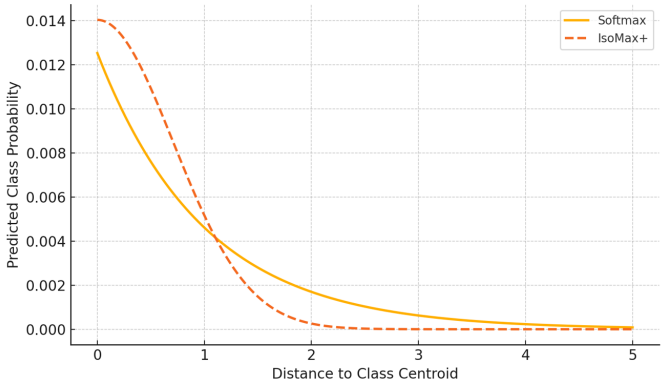


Fig. 1. Comparison of softmax and IsoMax+ confidence decay with respect to increasing feature distance from class centroids.

IsoMax+ loss overcome this limitation by using distance based approach instead of softmax function. This replacement limits marking unfamiliar data by attach for them lower confidence scores and high predictive entropy. In addition, by enhancing isotropic feature, this model is easily detect OOD data without requiring external datasets. Therefore, it is suitable with high dimension and noisy data. To clarify comparison between Softmax and IsoMax+, Fig. 1 visualizes how each method predict class based on distance of an input sample form the class centroid.

1) Softmax: The standard softmax computes class probabilities using linear logits:

$$p_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \tag{2}$$

where, z_i is typically computed as a linear projection: $z_i = w_i^\top f(x) + b_i$, with $f(x)$ being the input feature vector. This formulation does not explicitly depend on the distance between the input and any class centroid, which can lead to overconfident predictions for unfamiliar inputs (i.e., OOD data).

2) IsoMax+: In contrast, IsoMax+ replaces logits with negative distances in the embedding space:

$$p_i = \frac{e^{-d(f(x),c_i)}}{\sum_{j=1}^{K} e^{-d(f(x),c_j)}} \tag{3}$$

with the distance metric defined as:

$$d(f(x), c_i) = \|f(x) - c_i\|^2 \tag{4}$$

Here, c_i is the centroid of class i. This formulation ensures that the farther an input lies from any class center, the lower its predicted probability, thereby naturally increasing predictive entropy for OOD samples.

Recently, as the limitaion of Sofmax based, IsoMax+ which a part of EOOD framework addressed the entropy detection help model more practical when using for real world data.

III. Methodology

This section presents the EOOD framework for retail time-series. The pipeline has four stages: 1) data preprocessing and surrogate OOD labeling, 2) spectrogram transformation, 3) EfficientNet-B3 backbone with IsoMax+ loss, and 4) entropy-based OOD scoring.

A. Overview of the Proposed Framework

Fig. 2 summarizes the workflow. Raw transaction logs are aggregated into monthly series, transformed into spectrogram images via STFT, and then fed to an EfficientNet-B3 backbone trained with IsoMax+ loss for uncertainty-aware classification. The predictive distributions are finally converted into OOD scores.

B. Data Pre-processing and Labeling

We use transaction-level sales data from 63 provinces in Vietnam (2024), including product ID, date, quantity, revenue, and location. Records are aggregated monthly per (product, province) to form 12-month sequences; only pairs with complete records and at least six non-zero months are retained to avoid degenerate or overly sparse series.

To emulate realistic distributional shifts without relying on external OOD datasets, two complementary heuristics were applied for labeling, as illustrated in Fig. 4.

1) Spike-based labeling rule:: A month t is labeled as out-of-distribution (OOD) if its revenue R_t satisfies:

$$R_t > 2 \times R_{t-1}, \tag{5}$$

capturing short-term shocks such as promotions, sudden demand surges, or supply disruptions.

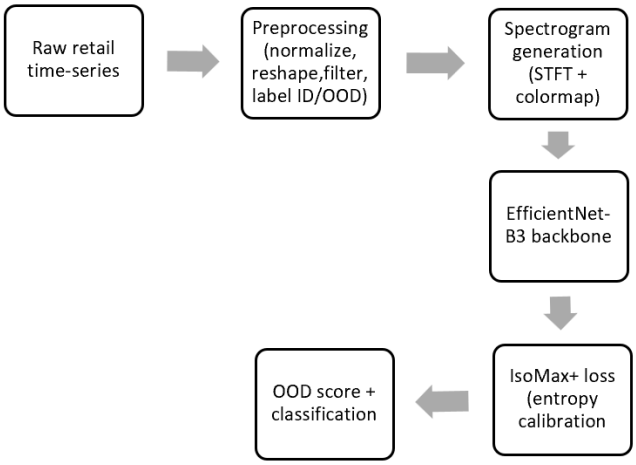


Fig. 2. Proposed EOOD framework for retail time-series OOD detection.

2) Entropy-based volatility labeling:: For each 12-month revenue series,

$\mathbf{R} = \{R_1, R_2, \ldots, R_{12}\}$,

we compute the volatility ratio (serving as an entropy proxy) as:

$$E = \frac{\sigma(\mathbf{R})}{\mu(\mathbf{R})}, \tag{6}$$

where, $\sigma(\mathbf{R})$ and $\mu(\mathbf{R})$ denote the standard deviation and mean of the sequence, respectively. A sequence is labeled as OOD if its entropy ratio exceeds a global threshold:

$$E > \mu_E + 1.5\sigma_E, \tag{7}$$

where, μ_E and σ_E represent the mean and standard deviation of E across all product–province pairs. This criterion identifies long-term volatility or regime shifts in sales behavior.

After pre-processing, a total of 33,680 valid product–province time-series pairs were obtained, including 22,450 in-distribution (ID) and 11,230 OOD sequences. These labeled sequences were subsequently transformed into spectrogram images for model training in the following section.

C. Spectrogram Transformation

Each 12-month sequence is converted into a spectrogram using the Short-Time Fourier Transform (STFT),

$$S(t, f) = \sum_{m=-\infty}^{\infty} x[m]\, w[m-t]\, e^{-j2\pi f m} \tag{8}$$

and mapped to a three-channel image via a fixed colormap. This representation preserves temporal order while exposing time–frequency structure, where retail anomalies often appear as localized spectral distortions. Prior studies show that such image-based encodings can improve time-series classification and anomaly detection in complex settings [6], [19]. Visual examples are shown in Section IV.

D. Model Architecture

We adopt EfficientNet-B3 as the backbone and resize spectrograms to 300×300 pixels. EfficientNet's compound scaling achieves a favorable accuracy–latency balance compared to heavier backbones, which is crucial for near real-time retail inference [20]. Global average pooling is applied to the final feature maps, and the ImageNet classifier is replaced by an IsoMax+ head.

E. IsoMax+ Loss for OOD Detection

Traditional softmax classifiers often produce overconfident predictions even for inputs that are far from the training distribution. This limitation arises because the Softmax function converts unnormalized logits into probabilities without explicitly considering the distance of a sample from any known class centroid. As a result, OOD samples can still be assigned high confidence scores, leading to poor uncertainty calibration [7].

IsoMax+ addresses this problem by replacing linear logits with distance-based formulations. Specifically, the logit for class i is computed as:

$$z_i = -\alpha \cdot \|\mathbf{f} - \mathbf{c}_i\|^2 \tag{9}$$

where, $\mathbf{f}$ is the feature vector of an input, $\mathbf{c}_i$ is the centroid of class i, and α is a learnable scaling parameter. This formulation ensures that samples located far from all class centroids are assigned low confidence and high predictive entropy, thereby mitigating the overconfidence issue without requiring post-hoc calibration [10].

To provide intuition, Fig. 3 compares the posterior probability distributions of Softmax and IsoMax+. As shown, Softmax tends to output peaked, low-entropy posteriors even for ambiguous inputs, while IsoMax+ yields smoother, high-entropy distributions that better reflect uncertainty. This property makes IsoMax+ particularly suitable for OOD detection in dynamic retail environments.

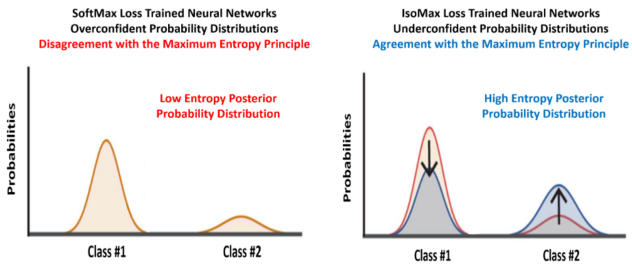


Fig. 3. Comparison of posterior probability distributions produced by Softmax and IsoMax+.

F. Training Strategy

Models are trained end-to-end with AdamW and a cosine learning rate schedule. We use random cropping, horizontal flips, brightness/contrast jitter, Gaussian noise, MixUp (α=0.4), label smoothing (0.1), and Exponential Moving Average of weights. These choices improve robustness to small rendering differences in spectrograms, temper overconfident posteriors, and stabilize convergence under volatile demand patterns typical of retail data.

In summary, the EOOD framework combines dual-criteria labeling, spectrogram-based encodings, an EfficientNet-B3 backbone, and IsoMax+ loss to deliver calibration-free OOD detection suited to dynamic retail environments.

IV. Experiments

A. Dataset Collection and Pre-Processing

The raw dataset contained approximately 1.2 million transaction records from 63 provinces in Vietnam, covering the 12-month period of 2024. The data were obtained from the retail sales management system of *Liberico Company*, which operates on the nationwide VNPost postal distribution channel. This channel supports multiple vendors and retail partners beyond Liberico itself, making it a rich and heterogeneous retail environment that includes diverse product categories, pricing schemes, and sales behaviors from different businesses. The dataset was directly extracted from the company's database and exported into excel format Table II summarizes the key attributes contained in the raw transaction dataset. After monthly aggregation and filtering for completeness (i.e., retaining only product–province pairs with at least six non-zero months), a total of 33,680 valid time-series sequences were preserved for analysis. These sequences were subsequently labeled based on two criteria:

- Spike-based labeling: A time-series is labeled OOD if any month shows a revenue spike exceeding 2× the previous month.
- Entropy-based labeling: The entropy of each sequence was computed as the ratio $\frac{\sigma}{\mu}$ (standard deviation over mean). If this value exceeded a global threshold of $(\mu + 1.5\sigma)$, the sequence was labeled as OOD.

TABLE II. Data Dictionary of the Retail Transaction Dataset

Field	Description
Date	Transaction date (dd/mm/yyyy), used to derive month.
Province	Province or city where the transaction occurred.
Invoice ID	Unique identifier of the invoice.
Product ID	Unique identifier of the product.
Product Name	Description of the product.
Unit	Measurement unit (e.g., piece, pack, bottle).
Unit Price	Price per unit (VND).
Quantity	Number of units sold.
Revenue	Total transaction revenue (Unit Price × Quantity).
Commission Rate	Commission ratio applied to the transaction.
Commission (derived)	Calculated as Revenue × Commission Rate.
Total Payment (derived)	Calculated as Revenue − Commission.
Month (derived)	Extracted from the Date for time-series grouping.

In total, 22,450 sequences were labeled ID, and 11,230 were labeled OOD. The distribution of entropy values for ID and OOD sequences is shown in Fig. 4.

Each of the 33,680 time-series sequences was converted into a spectrogram using the Short-Time Fourier Transform. The resulting spectrograms were mapped to a 3-channel image using a fixed colormap. This conversion produced a dataset

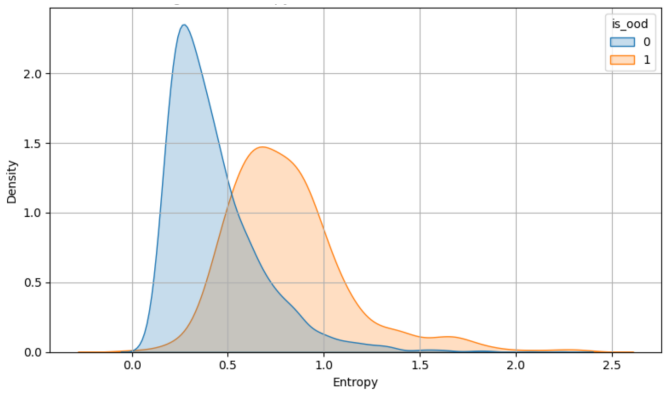


Fig. 4. Distribution of entropy scores for labeling ID (blue) vs. OOD (orange) sequences.

of 33,680 spectrogram images, each of size 300×300 pixels, which were then organized into two directories: one for ID (label 0) and one for OOD (label 1). An example spectrogram of a 12-month revenue time series is shown in Fig. 5.

To improve generalization and avoid overfitting, we applied several data augmentation techniques as followings:

- Random horizontal flip: This was used to mimic random variations in input patterns (e.g., seasonal variations).
- Random cropping: Spectrogram images were randomly cropped to 270×270 pixels during training. This allowed the model to focus on localized features within the image, improving robustness.
- Brightness jitter: Slight changes to the brightness of the images were made to mimic changes in lighting conditions in the retail environment.
- Gaussian noise injection: Small amounts of noise were added to the images to simulate real-world signal fluctuations.
- MixUp ($\alpha = 0.4$): This technique mixes two random images during training, allowing the model to generalize better by learning from both images at once, which also increases the model's robustness against noisy data.

Fig. 5. Example spectrogram of a 12-month revenue series.

B. Experimental Setup

1) Training parameters: Following the methodology pipeline (Fig. 2), the spectrogram images were used as input to an EfficientNet-B3 backbone pre-trained on ImageNet. The model was trained with IsoMax+ loss for out-of-distribution (OOD) detection. We used the AdamW optimizer with an initial learning rate of 2×10^{-5} and a cosine-annealing schedule that progressively reduced the learning rate to ensure stable convergence. The loss function and learning rate scheduler were selected to balance optimization stability and convergence speed.

2) Exponential Moving Average (EMA): To further stabilize training and ensure that the model's weights converge to a stable point, we applied the Exponential Moving Average (EMA) of model weights. This technique averages the weights over time to ensure that fluctuations in model parameters are minimized, resulting in a more robust model that performs better in real-world scenarios [21]. The decay factor was set to 0.995, meaning that each new weight update is blended with the previous one by 99.5%.

3) Training environment: Training was carried out on NVIDIA Tesla T4 GPU with a batch size of 32. This GPU is highly efficient for training models like EfficientNet and provides a good balance between memory and processing power, making it suitable for high-throughput deep learning tasks. The model was trained for 75 epochs, and the early stopping criterion was applied, which halted training if the validation loss did not improve for 10 consecutive epochs. This approach helped prevent overfitting and ensured that training time was efficiently managed.

The model performance was evaluated based on classification accuracy, F1-score, AUC-ROC, and inference time(measured in milliseconds per sample).

C. Results

1) Phase 1: Baseline training and convergence analysis: Fig. 6 shows the accuracy and loss curves during the first phase of training with an 80/20 split. The model achieved rapid convergence within the first 15 epochs, stabilizing around 89% accuracy, while the validation loss decreased and plateaued at approximately 0.37. The confusion matrix in the figure shows that 48 out of 71 OOD samples were correctly classified, while 23 were misclassified as in-distribution. Table III shows the classification report for Phase 1.

TABLE III. CLASSIFICATION REPORT FOR PHASE 1

Class	Precision	Recall	F1-Score
In-Distribution	0.94	0.93	0.93
Out-of-Distribution	0.63	0.68	0.65
Accuracy		0.89	
Macro avg	0.79	0.80	0.79
Weighted avg	0.89	0.89	0.89

2) Phase 2: Enhanced training and robustness evaluation: In the second phase, the dataset was split into 70% for training, 15% for validation, and 15% for testing. This phase included advanced regularization techniques, such as MixUp, label smoothing, and EMA, to improve generalization and

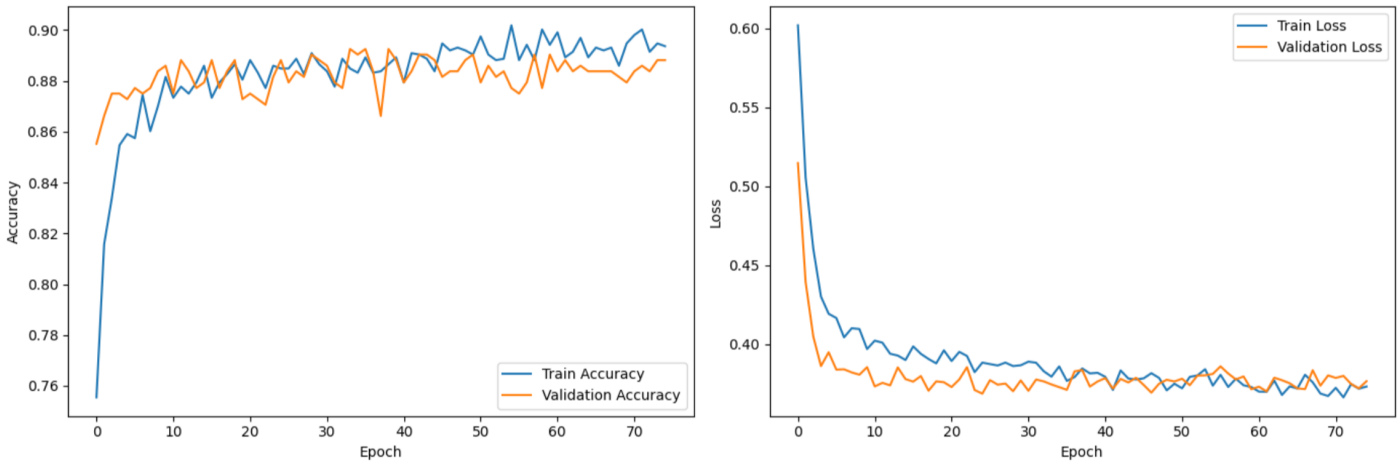


Fig. 6. Accuracy and loss curves during Phase 1.

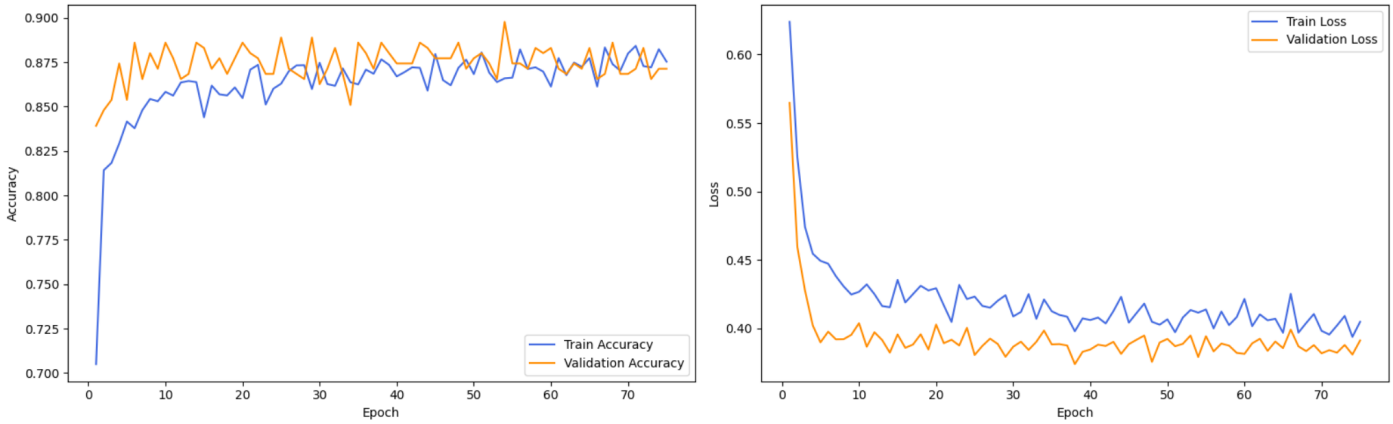


Fig. 7. Accuracy and loss curves during Phase 2.

reduce overfitting. The model performed consistently well, with training and validation accuracy stabilizing 87% after epoch 30, as shown in Fig. 7. The validation loss remained slightly lower than the training loss throughout the phase, indicating stable generalization.

TABLE IV. CLASSIFICATION REPORT FOR PHASE 2

Class	Precision	Recall	F1-Score
In-Distribution	0.91	0.92	0.92
Out-of-Distribution	0.69	0.65	0.67
Accuracy		0.87	
Macro avg	0.80	0.79	0.79
Weighted avg	0.86	0.87	0.86

The detailed classification performance is summarized in Table IV, where the model achieved an overall accuracy of 87%. The In-Distribution class obtained high precision and recall (0.91 and 0.92, respectively), while the Out-of-Distribution class was more challenging, with a precision of 0.69 and recall of 0.65. These results highlight the model's improved robustness compared to Phase 1, although OOD detection remains the more difficult task.

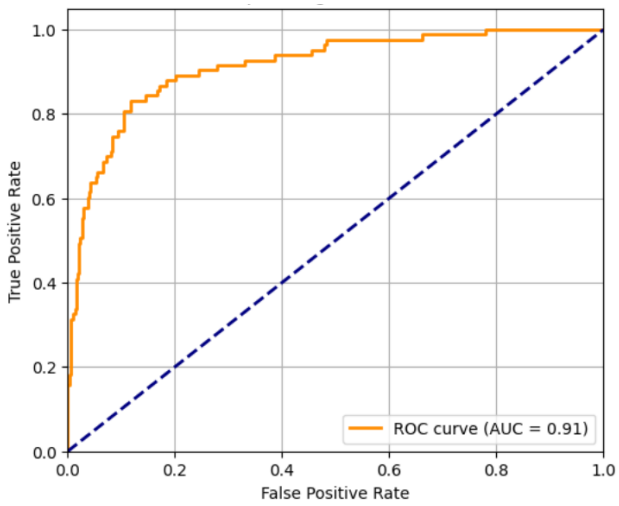
Additionally, Fig. 8 presents the ROC curve of the best EMA-tracked model evaluated on the 15% test set. The model achieved an AUC of 0.7974, demonstrating its strong ability to distinguish between in-distribution and OOD samples.

3) Comparison with baseline methods: To comprehensively evaluate the performance of the EOOD framework, we compared it with several baseline OOD detection methods: MSP, Entropy Scoring, ODIN, and Mahalanobis Distance. Table V shows the comparison results based on key metrics: F1-score (OOD), AUC–ROC, and inference time.

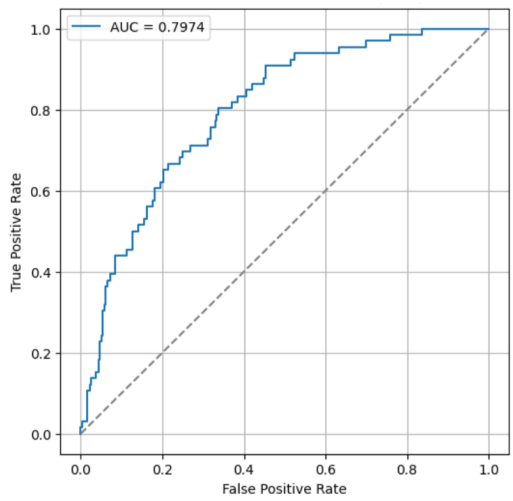
EOOD consistently outperforms all baseline methods across the key metrics. It achieves the highest F1-score of 0.88 for OOD detection and the best AUC–ROC of 0.91, demonstrating superior ability to distinguish between ID and OOD samples. In contrast, while ODIN and Mahalanobis show competitive AUCs, their inference times are significantly higher, making them less suitable for real-time deployment. Methods like MSP and entropy scoring suffer from low OOD recall due to overconfidence in softmax outputs.

D. Discussion

The experimental findings presented in Section IV demonstrate that the proposed EOOD framework effectively enhances out-of-distribution (OOD) detection for retail time-series data.



(a) ROC Curve on 80/20 split (Phase 1).



(b) ROC Curve on 70/15/15 split (Phase 2).

Fig. 8. Comparison of ROC curves across two training configurations.

TABLE V. PERFORMANCE COMPARISON BETWEEN EOOD AND TRADITIONAL OOD DETECTION METHODS, SORTED BY F1-SCORE

Method	**F1-score (OOD)**	**AUC–ROC**	**Inference Time** (ms)
MSP	0.22	0.80	9
Entropy	0.36	0.80	9
Energy	0.41	0.75	11
Mahalanobis	0.81	0.86	17
ODIN	0.83	0.89	19
EOOD	**0.88**	**0.91**	**12**

By transforming sales sequences into spectrogram representations and employing the IsoMax+ loss, the model improves uncertainty calibration and maintains efficient inference suitable for real-time applications.

From the comparative experiments, the Phase 1 baseline achieved 0.89 accuracy and an F1-score of 0.65 for OOD samples, whereas the enhanced Phase 2 configuration—with MixUp, Label Smoothing, and Exponential Moving Average (EMA)—achieved smoother convergence and more stable generalization. Although the accuracy slightly decreased to 0.87, the model exhibited lower variance in validation loss and better calibration, as seen in Fig. 7. This trade-off indicates improved robustness and reduced overfitting under dynamic retail conditions.

As shown in Table V, EOOD outperformed Mahalanobis and ODIN, achieving the highest overall accuracy and efficiency. These results verify that the entropic–distance synergy inherent in IsoMax+ enables superior calibration-free OOD detection compared to confidence-based and post-hoc methods. In particular, EOOD achieved an approximate 5% gain in AUC over Mahalanobis Distance while reducing latency by nearly 35%, offering a practical trade-off between accuracy and computational cost.

Despite these promising results, Fig. 8(b) also shows that the recall for OOD samples remains modest (0.65), implying that extreme or rare anomalies may still overlap with in-distribution variance. This suggests the need for adaptive thresholding or dynamic decision boundaries to further improve sensitivity under evolving retail conditions.

The framework's scalability and calibration-free operation make it well-suited for deployment in production environments. Beyond retail, similar entropic representations could be applied to other time-series domains such as financial fraud detection, healthcare monitoring, or industrial sensor analysis, where rapid and reliable anomaly identification is critical.

The framework's scalability and calibration-free operation make it well-suited for deployment in production environments. Beyond retail forecasting, the proposed framework can also support anomaly detection tasks in finance, healthcare, and industrial monitoring, where timely and reliable OOD identification is critical. Its calibration-free nature and low inference latency make it practical for large-scale real-time deployment.

Overall, these findings confirm that entropy-based learning combined with distance-aware regularization forms a strong foundation for robust, scalable, and interpretable OOD detection across complex and dynamic business environments.

V. CONCLUSION

This study presented an enhanced Entropic Out-of-Distribution Detection (EOOD) framework specifically designed for dynamic retail time-series data. The proposed approach integrates IsoMax+ loss with spectrogram-based representation learning to improve robustness against distributional shifts — a persistent challenge in real-world retail environments. By transforming sales sequences into visual spectrograms, the model effectively captures temporal dependencies and subtle anomalies in a spatial format, extending the applicability of convolutional architectures beyond traditional image domains.

Unlike conventional OOD techniques that require external datasets or post-hoc calibration, the proposed framework operates in a fully calibration-free manner, making it practical for business applications where real-time decisions are critical. This design promotes scalability and adaptability across domains with high-dimensional, sparse, or unlabeled data, supporting broader applications in predictive modeling and anomaly detection.

The findings demonstrate that entropy-based learning, when combined with distance-aware regularization, can significantly enhance uncertainty calibration and detection accuracy. The framework offers both scientific and practical benefits, contributing to more reliable forecasting and decision-making in data-driven retail operations.

A. Limitations

Despite its strong empirical performance, the proposed EOOD framework still has several limitations. First, the labeling process for in-distribution and out-of-distribution samples relies on heuristic thresholds (e.g., entropy ratio and revenue spike rules), which may not generalize optimally across different datasets or business domains. Second, the current approach focuses on short-term (12-month) retail patterns, while longer seasonal or multi-year dynamics have not been explored. Finally, converting temporal sequences into spectrograms, though effective for CNN-based learning, may introduce minor information loss compared to direct temporal modeling approaches such as Transformers or recurrent architectures. These limitations highlight opportunities for improvement in generalization and interpretability.

B. Future Work

Several directions can be pursued to extend this research. First, adaptive thresholding techniques could further enhance sensitivity to gradual concept drift in long-term deployments. Second, integrating Bayesian-inspired uncertainty modeling may strengthen interpretability and improve predictive calibration. Finally, cross-domain evaluation in fields such as finance, cybersecurity, and energy analytics could validate the generalizability of the proposed framework and highlight its potential for broader adoption.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58, 2009.

[2] Z. Chen and B. Liu, *Lifelong Machine Learning*. Morgan & Claypool Publishers, 2018.

[3] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 7167–7177, 2018.

[4] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[5] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014.

[6] Y. Zhang and et al., "Entropy-based out-of-distribution detection," *Journal of Machine Learning Research*, vol. 22, no. 10, pp. 1234–1256, 2021.

[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 1321–1330.

[8] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset Shift in Machine Learning*. MIT Press, 2009.

[9] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[10] D. Macêdo, T. I. Ren, C. Zanchettin, A. L. Oliveira, and T. Ludermir, "Entropic out-of-distribution detection," in *International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021.

[11] S. Liang, C. Wang, and Y. Zheng, "Entropy regularization for reliable out-of-distribution detection in deep networks," *Neural Networks*, vol. 173, p. 106208, 2024.

[12] Y. Han, H. Zhang, and B. Li, "Ood detection for time-series via contrastive predictive entropy," *Pattern Recognition Letters*, vol. 173, pp. 13–21, 2023.

[13] Q. Liu and et al., "Application of neural network-based out-of-distribution detection in healthcare systems," *Journal of Healthcare Systems Research*, vol. 34, no. 3, pp. 245–259, 2022.

[14] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[15] Y. Liu, X. Wang, and Y. Jin, "Improving out-of-distribution detection through spectral representation learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 11–23, 2023.

[16] D. Macedo, T. I. Ren, and C. Zanchettin, "Isomax+ and beyond: Revisiting distance-aware classifiers for robust out-of-distribution detection," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2023.

[17] S. Yang, J. Li, X. Liu, and J. Xu, "A survey on out-of-distribution detection in deep learning: Taxonomy, challenges, and future directions," *Information Fusion*, vol. 103, p. 102070, 2024.

[18] Y. Jin and et al., "A study on the effectiveness of entropy for detecting out-of-distribution data," *Journal of Machine Learning Research*, vol. 23, no. 5, pp. 987–1003, 2022.

[19] N. Hatami, Y. Gavet, and J. Debayle, "Classification of time-series images using deep convolutional neural networks," in *International Conference on Machine Vision (ICMV)*. SPIE, 2017, p. 87790A.

[20] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 2019, pp. 6105–6114.

[21] P. Izmailov, D. Molchanov, T. Garipov, C. Blundell, and D. Vetrov, "Averaging weights leads to wider optima and better generalization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 4300–4310.

A Review of Artificial Intelligence in Inventory Management: Methods, Applications and Directions

Jinjin Li[1], Huijun Huang[2]*, Yuping Gong[3], Lei Wang[4], Xiangui Yin[5], Yichang Liu[6]
Metering Center, Guangxi Power Grid Co., Nanning, China[1,2,3,4]
Inspur Group, Jinan, China[5,6]

***Abstract*—Effective inventory management is fundamental to supply chain resilience and efficiency. Artificial intelligence (AI) has emerged as a transformative solution that enables more dynamic and data-driven inventory strategies. To map the latest advancements in this rapidly evolving field, this study presents a systematic literature review (SLR) of AI techniques in inventory management. The review was conducted following the PRISMA 2020 guidelines, through which 87 high-quality articles published between 2021 and 2025 were systematically analyzed. Our review identifies machine learning (ML), deep learning (DL), reinforcement learning (RL), and hybrid methods as the predominant AI technologies. These techniques primarily address three foundational tasks. In demand forecasting, they improve prediction accuracy and mitigate stockout and overstock risks. For inventory control, they balance costs with service levels and optimize replenishment strategies. In inventory classification, they facilitate targeted resource allocation. Despite these advancements, AI research confronts significant challenges, particularly in data dependency, model interpretability, and implementation overhead. To address these gaps, we suggest future research focused on data-efficient learning, explainable AI, and lightweight, integrated frameworks to lower adoption barriers. This review provides a timely and holistic overview of the current research landscape, which serves as a reference for academics to identify research directions.**

Keywords*—Inventory management; artificial intelligence; demand forecasting; inventory control; inventory classification; machine learning

I. Introduction

Effective supply chain management (SCM) plays a critical role in modern corporate strategy by optimizing resource allocation, enhancing market responsiveness, and mitigating operational risks [1], [2]. Inventory management is a core component of SCM that is crucial for balancing supply and demand. It directly impacts key performance indicators such as operational costs and customer satisfaction [3], [4]. The strategic importance of inventory management is particularly evident in high-stakes sectors, such as the power grid sector. In this domain, managing inventories of critical assets, such as transmission components and essential spare parts, is paramount for ensuring grid stability and operational reliability, especially amidst the growing integration of volatile renewable energy sources [5], [6]. Traditional inventory management methods, reliant on static models and historical data, are ill-equipped for the volatility of modern supply chains [7], [8]. These methods exhibit inherent limitations in timeliness, flexibility, and predictive accuracy. Consequently, there is a growing demand for advanced solutions, and AI is a particularly promising approach [9], [10].

AI leverages big data and intelligent algorithms to enable scientific forecasting and dynamic optimization in inventory management, improving efficiency, reducing costs, and enhancing supply chain flexibility [11], [12]. For instance, machine learning (ML) can perform demand forecasting based on historical sales and market trends [13]. Deep learning (DL) is capable of processing complex nonlinear data for anomaly detection and defective product identification [14]. Natural language processing (NLP) can analyze supplier and customer information to support market fluctuation prediction [15]. Reinforcement learning (RL) can optimize replenishment strategies in dynamic environments [16]. The Internet of Things (IoT) combined with computer vision (CV) enables real-time monitoring, automated stocktaking, and visualization management [17]. These techniques significantly improve forecasting accuracy, inventory turnover, and anomaly detection, while supporting strategic supply chain planning and promoting intelligent manufacturing and digitalized supply chains [18][19][20].

Although artificial intelligence is rapidly transforming inventory management, the field lacks a current, unified map of these advancements, creating a research gap. To the best of our knowledge, there is a notable scarcity of comprehensive systematic reviews on this topic published after 2023. This gap is particularly critical for two reasons. On one hand, recent years have witnessed unprecedented progress in AI, with the emergence of large language models (LLMs) and the maturation of deep reinforcement learning and Transformers opening new paradigms. On the other hand, existing research has become increasingly fragmented. Studies often focus on narrow problems, such as demand forecasting or stock control, using isolated techniques, without providing a holistic view. Consequently, a systematic synthesis is urgently needed to answer critical questions regarding the current research landscape, the primary application topics, the comparative advantages and limitations of different AI techniques, and the challenges and future directions.

To address this clearly defined gap, this study presents a systematic literature review of AI in inventory management focused on the crucial 2021 to 2025 period. By synthesizing the latest advancements, we provide an updated and holistic framework of the current research status, topics, techniques, and future directions. This study, therefore, offers significant contributions. For the academic community, it maps the fragmented research landscape, identifies critical knowledge

*Corresponding author.

gaps, and outlines promising avenues for future inquiry. For practitioners and industry leaders, it serves as a practical guide, demystifying the capabilities and limitations of various AI methods and offering evidence-based insights to support strategic technology adoption. Ultimately, this review acts as a timely and essential reference, empowering researchers to navigate the new frontier and enabling practitioners to implement state-of-the-art solutions that enhance supply chain resilience and efficiency.

The remainder of this study is structured as follows: Section II introduces the research methodology employed for this review. Section III presents a descriptive analysis of the collected literature over the past five years. Section IV provides a detailed technical analysis based on the literature review. Section V discusses the findings in relation to the key research questions posed in this study. Finally, Section VI concludes the study.

II. Research Methodology

Due to its rigor, transparency, and reproducibility, the systematic literature review (SLR) enables a comprehensive and evidence-based synthesis of AI techniques and applications in inventory management. To further enhance the transparency and scientific rigor of this study, the entire process follows the preferred reporting items for systematic reviews and meta-analyses (PRISMA 2020) guidelines as shown in Fig. 1. The research began by outlining four research questions to explore the state, main topics, applied methods, and future directions of AI in inventory management. Subsequently, literature searches were conducted across databases. The retrieved articles were then filtered according to specific inclusion criteria. Studies unrelated to the topic, non-academic sources, and those outside the publication window were excluded. Finally, the selected literature was systematically analyzed to identify major findings, research trends, challenges, and recommendations for future research.

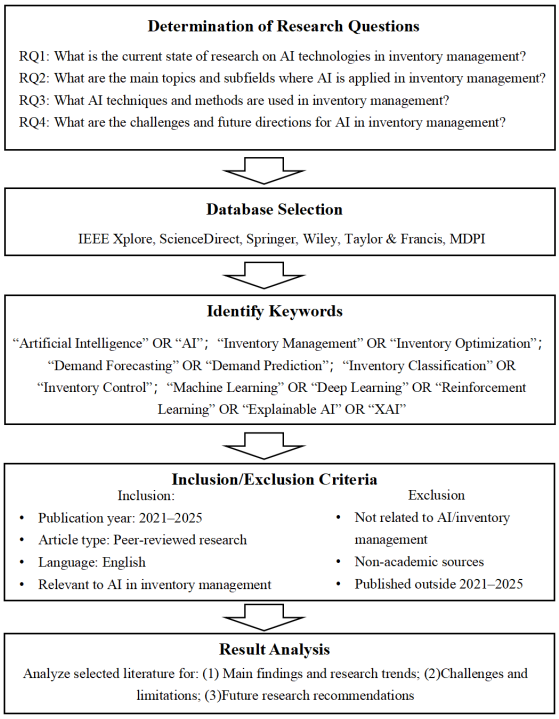


Fig. 1. Flowchart of the proposed method.

A. Research Questions

In the context of AI applications in inventory management, this study addresses the following research questions:

RQ1: What is the current state of research on AI technologies in inventory management?

RQ2: What are the primary topics in inventory management that AI is applied to solve?

RQ3: What are the key AI techniques and algorithms applied in inventory management, and what are their respective advantages and disadvantages?

RQ4: What challenges does AI face in inventory management, and what directions can be explored in future research?

B. Literature Search Strategy

To ensure comprehensive coverage of this interdisciplinary topic, our literature search utilized several major academic databases. The chosen platforms, IEEE Xplore, ScienceDirect, Springer, Wiley, Taylor & Francis, and MDPI, were selected as they collectively represent the primary, high-quality publication venues for research spanning artificial intelligence, engineering, and supply chain management. This multi-database strategy was crucial for minimizing selection bias and capturing a broad spectrum of relevant studies.

A combination of subject terms and free-text keywords are used: ('Artificial Intelligence' OR 'AI') AND ('Inventory Management' OR 'Inventory Optimization') AND ('Demand Forecasting' OR 'Demand Prediction' OR 'Inventory Classification' OR 'Inventory Control') AND ('Machine Learning' OR 'Deep Learning' OR 'Reinforcement Learning' OR 'Explainable AI' OR 'XAI'). Only English-language papers were considered. After initial screening, a total of 3,085 publications related to the above topics were identified.

C. Inclusion and Exclusion Criteria

The inclusion criteria required that studies focus on the application or methodology of AI in inventory management and related domains, covering techniques such as ML, DL, RL and other related methods. Only peer-reviewed research papers published between 2021 and 2025, with sufficient methodological or practical details, were included. Studies unrelated to AI in inventory management, non-academic publications, or published outside the specified period were excluded.

D. Quality Assessment

To ensure the scientific rigor and reliability of the systematic literature review, all included studies were subjected to a structured quality assessment based on established SLR frameworks [21]. The assessment focused on the clarity of research objectives, the appropriateness of the AI techniques employed, the transparency of data and experimental procedures,

the completeness and reliability of results, and the practical relevance of the study. Finally, 87 publications remained for further research.

III. Results

Following the research protocol described above, the literature relevant to the topic was selected and analyzed from four perspectives: publication trends, resource distributions, most cited papers, and keyword analysis.

A. Publication Trend

Fig. 2 presents the publication trend in the field over the past five years using a line chart. Overall, the number of publications increased significantly from 6 in 2021 to 22 in 2025, with a particularly notable surge in 2022, reflecting the rapid growth of academic interest in this topic. The number of publications remained relatively stable in the subsequent years, indicating sustained attention from the research community.

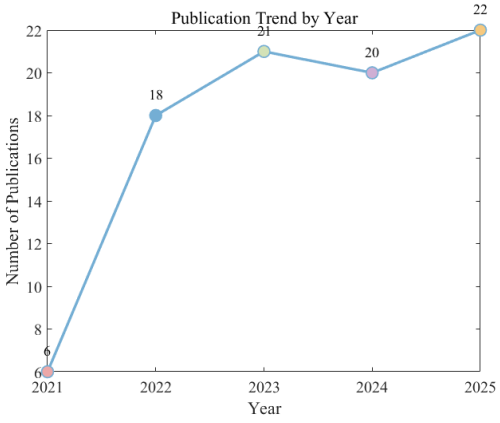


Fig. 2. Number of publications by year.

B. Resource Distributions

The distribution of publications across academic databases reflects the accessibility and concentration of research in this field. As shown in Fig. 3, *IEEE Xplore* hosts the largest number of publications (24 papers), followed by *ScienceDirect* (21 papers) and *Springer* (16 papers). This uneven distribution indicates that *IEEE Xplore*, *ScienceDirect*, and *Springer* serve as the primary sources of relevant studies, while *MDPI*, *Wiley*, *Taylor & Francis* and *Others* play a supplementary role.

At the journal level, the selected publications are relatively concentrated. Among the source journals, *Applied Sciences* contributed the most articles (8 papers), followed by *IEEE Access* with 7 papers. Additionally, *Annals of Operations Research* and *International Journal of Production Research* each published 4 relevant articles. Other contributing journals include *Sustainability* (3 papers), *Journal of Intelligent Manufacturing* (2 papers), *Expert Systems with Applications* (2 papers), *European Journal of Operational Research* (2 papers), *Production and Operations Management* (2 papers), and *IEEE Transactions on Engineering Management* (2 papers).

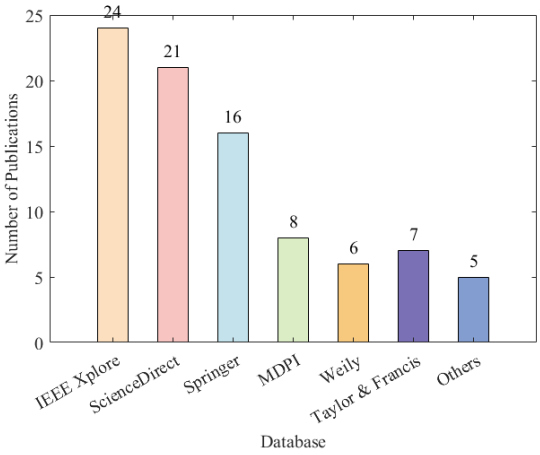


Fig. 3. Distribution of screened publications across databases.

C. Most Cited Papers

To illustrate the recent activity in this research area, Table I lists the top ten most cited papers. Leading publications received over 100 citations within a short period. For instance, the paper by Moor [22] has 126 citations, and the paper by Gonccalves [23] has 115 citations. The high citation counts of recent publications further demonstrate the active research in this field.

TABLE I. Top 10 Most Cited Papers

Rank	Authors	Reference	Citations
1	Moor et al.	[22]	126
2	Gonccalves et al.	[23]	115
3	Wang et al.	[24]	114
4	Meisheri et al.	[25]	111
5	Abu Zwaida et al.	[26]	109
6	Selukar et al.	[27]	94
7	Tang et al.	[28]	85
8	Jahin et al.	[29]	85
9	Shajalal et al.	[30]	79
10	Xu et al.	[31]	78

D. Keyword Analysis

As shown in Fig. 4, the keyword frequency analysis indicates that research in this field primarily focuses on the application of AI in inventory management. 'Inventory management' (21 occurrences) appears most frequently, followed by key methodological terms such as 'machine learning' (15 occurrences), 'deep learning' (10 occurrences), and 'reinforcement learning' (8 occurrences). Additionally, high-frequency terms like 'inventory control' (13 occurrences) and 'demand forecasting' (12 occurrences) further confirm that AI techniques are actively applied to address core challenges in this domain.

IV. Review of Literature

This section provides a comprehensive review of the core literature identified through a systematic search, with a focus on the AI techniques in inventory management. To clearly present the research landscape, the studies are categorized according to the main tasks of inventory management, including demand forecasting, inventory control, inventory classification,

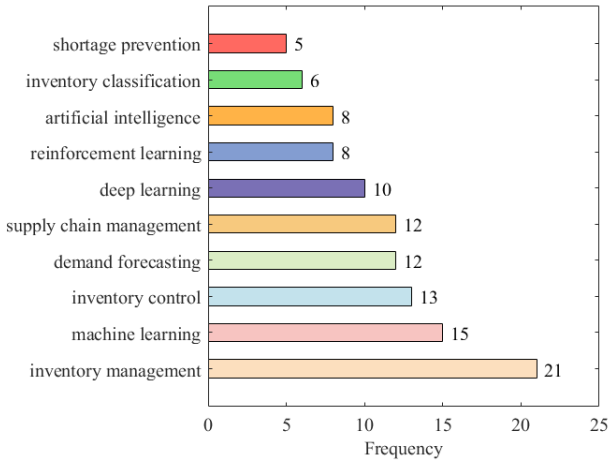


Fig. 4. Top 10 keywords by frequency.

and other related topics and application examples. Within each topic, the methods and algorithms such as ML and DL are reviewed. As this section contains a substantial number of technical acronyms, a table of abbreviations and their full names is provided in Appendix A for easy reference [see Table V].

A. Demand Forecasting

Accurate demand forecasting is central to inventory management, as it guides stock level decisions and mitigates the risks of stockouts and overstocking. A summary of works in this area is provided in Table II. Furthermore, demand forecasting can be broken down into more specific tasks, such as demand forecasting, inventory level prediction, and sales and price prediction.

1) Demand forecasting: Demand forecasting represents a core task in inventory management, aiming to predict future product or service demand based on historical data and relevant influencing factors. In existing research, mainstream approaches primarily encompass ML and DL techniques.

ML provides a simple and efficient way to predict the demands. Sathish [40] compared Exponential Smoothing (ES) with Gradient Boosting (GB), and found that ES outperformed GB in both forecasting accuracy and inventory optimization. Gonccalves [23] compared ML methods with the statistic based ARIMAX, and proved that ML models achieve superior results in later stages. Manos [49] developed a dynamic demand forecasting platform with XGBoost and SVR to process the multiple data source and then improve the prediction accuracy. Yang [44] developed a dynamic forecasting platform that combines XGBoost and SVR, which leverages multi-source data to significantly enhance forecasting accuracy and operational efficiency.

DL models have shown remarkable success in supply chain forecasting, with predictive power scaling alongside model complexity. Foundational architectures like LSTM networks, have been widely adopted. For instance, Deng [32] utilized LSTMs to capture long-term temporal dependencies and achieved prediction accuracies exceeding 80%. The versatility of LSTMs is further demonstrated in specialized applications, such as the hybrid LSTM variant proposed by Hasni [33] for intermittent medical supply demand, and the integration with customer segmentation by Jauhar [39], which slashed prediction errors by 74.14%. Building on this foundation, more sophisticated models are proposed to further enhance forecasting accuracy. Liu [45] developed a BO-CNN-LSTM model where Bayesian optimization fine-tunes a combined CNN-LSTM architecture. Seyedan [50] employed an ensemble DL method to forecast online retail demand, and the predictions are further used to optimize safety stock. Jin [51] integrated an attention-based Temporal Convolutional Network (ATCN) into their system. Abualuroug [43] proposed the MMCW-BiLSTM model which leverages multi-dimensional collaborative attention. For industry-specific challenges, models like the NARXNN have also been effective, and the accuracy reaches 96.24% in food demand forecasting [41]. In addition, the Transformer architecture is also gaining traction. Yang et al. [44] combined a Transformer with multi-agent RL, and yielded an 18.2% reduction in prediction error.

2) Inventory level prediction: Predicting inventory levels and stockout risks is crucial for optimizing inventory control and minimizing costs. Researchers have applied a range of ML and DL models to this task.

As to ML methods, tree-based ensemble models like XGBoost have proven effective for direct inventory level prediction [28] and related tasks such as order quantity forecasting [47]. DL, on the other hand, provides more powerful tools to capture complex nonlinear patterns. For instance, a hybrid Multi-MLP and LightGBM model yielded a significantly lower prediction error with an MAE of 0.2331 [34]. Novel architectures target specific challenges: QAmplifyNet, a quantum-classical hybrid network, was designed to handle data imbalance in stockout prediction and achieved 90% accuracy [29]. To address the need for model transparency, an explainable CNN framework [30] was proposed, which delivered both high performance (AUC = 0.9489) and valuable interpretability.

3) Sales and price prediction: Accurate sales and price prediction is fundamental to developing effective commercial strategies, enabling businesses to forecast product demand and establish optimal pricing. Both ML and DL have been extensively applied to address these interconnected challenges.

For prediction tasks, ML models like SVR, MLR, and MNLR have demonstrated high accuracy of 95.63% [36], while more advanced hybrid kernel models such as IMOHHO-AHK-ELM have been developed for specialized areas like wine price forecasting [46]. DL architectures offer further improvements. BPNs have been used for daily sales prediction of perishable goods [35], and integrated CNN-LSTM models have proven effective for complex forecasting, leading to inventory cost savings of over 20% [38]. A real-time automatic pricing system was built using a Dueling DQN algorithm. This system [37] successfully tackled multi-period pricing challenges, securing 91.58% of the returns achievable by an optimal strategy.

B. Inventory Control

The central challenge in inventory control is to guarantee product availability while keeping total costs to a minimum.

TABLE II. LITERATURE FOR DEMAND FORECASTING

Study	Citations	Type	Algorithms	Method
[32]	78	DL	LSTM	Inventory management is optimized by using LSTM for demand prediction in this paper.
[33]	8	DL	LSTM	The challenge of accurately forecasting medical item demand in humanitarian operations is tackled using a hybrid LSTM method.
[23]	109	ML	VAR, ARIMAX, ANN	Multi-step demand forecasting in an automotive supply chain is tackled with a multivariate approach to improve prediction accuracy.
[34]	8	DL	MLP	Future inventory levels in manufacturing are predicted by employing a multi-MLP model.
[28]	69	ML	XGBoost	Inventory forecasting for cross-border e-commerce is enhanced using XGBoost to predict demand, with the aim of improving service automation and reducing costs.
[35]	13	DL	BPN	The daily sales quantity for short-shelf-life cream puffs is predicted using DL in this model, in order to reduce food waste.
[36]	79	ML	SVR, MLR, MNLR	The monthly demand for construction machinery spare parts (specifically manifold products) is predicted with ML and regression methods.
[37]	9	DL	DQN, DRL	The optimal real-time price for retailers in presale e-commerce environments was predicted using deep RL.
[38]	8	DL	DNN, CNN, LSTM	A dynamic sales prediction model (DSP-FAE) is proposed to forecast product sales volumes for inventory optimization, by which costs are reduced by over 20% using auto-learning and elastic-adjustment mechanisms.
[39]	15	DL, ML	K-means, Regression, LSTM	An explainable AI framework is proposed to predict wine sales volume (MntWines) for perishable supply chains by segmenting customers and forecasting demand per segment, resulting in a 74.14% reduction in forecasting errors.
[40]	115	DL	ES/GB	A material demand prediction task is solved to forecast material requirements through a comparative analysis of forecasting methods.
[41]	33	DL	NARXNN	A food demand prediction task is solved using a NARXNN model to forecast the daily sales volume of processed foods.
[42]	28	DL	LSTM	The inventory demand prediction task for volatile SME data is solved using LSTM and 2D kernel density estimation.
[43]	2	DL	BiLSTM	The inventory demand prediction task is solved using a novel MMCW-BiLSTM model, by which future product demand is predicted.
[44]	27	DL	MARL, Transformer, DQN	A multi-agent deep RL framework is proposed to solve demand prediction tasks in retail supply chains, achieving an 18.2% lower forecast error and 23.5% reduced stockout rates.
[45]	28	DL	CNN, LSTM	Demand forecasting is solved with BO-CNN-LSTM to predict demand accurately, which leads to improved inventory optimization and cost control.
[31]	14	DL, ML	LSTM, PSO	The demand prediction task in supply chains is solved using AI, specifically by forecasting future demand quantities to achieve higher accuracy and optimize resource allocation.
[46]	6	ML	IMOHHO-AHK-ELM	An IMOHHO-AHK-ELM prediction model is proposed to achieve accurate point and interval forecasting of wine prices using AI techniques.
[29]	34	ML	Quantum Neural Network	Supply chain backorders are predicted with 90% accuracy by QAmplifyNet using a hybrid quantum-classical neural network.
[47]	85	ML/DL	XGBoost, LSTM	The task of forecasting the number of meal orders for the next 10 weeks is addressed using time series analysis to minimize waste.
[30]	85	DL	CNN	A CNN-based explainable framework for product backorder prediction in inventory management was proposed in this paper.
[48]	-	DL	RNN, LSTM, GRU, CNN	A DL framework is proposed to predict supply chain risks; specifically, drug delivery status, food demand quantities, car prices, and electricity volumes are forecasted.

This involves a continuous balancing act between the costs of holding inventory, placing orders, and running out of stock. The primary tools for managing this balance are the optimization of order quantities, reorder points, and safety stock. A well-designed inventory system prevents stockouts, which directly improves customer satisfaction, operational efficiency, and overall profitability. In this section, we summarize the recent methods compiled in Table III. The task of inventory control can be further divided into three sub-categories: single-echelon inventory control, multi-echelon inventory control, and control strategies for special scenarios.

1) Single-echelon inventory control: The single-echelon inventory control task aims to determine the optimal replenishment policy for a single inventory node. Singh [62] combined decision trees with symbolic distance methods to defuzzify the fuzzy total cost of seasonal demand and imperfect perishable products, and the method achieved precise inventory control under uncertainty. Mo [64] proposed an online RL approach that dynamically adjusts inventory policies under non-stationary stochastic market demand without requiring prior knowledge of demand patterns.

2) Multi-echelon inventory control tasks: The multi-echelon supply chain inventory control task, which is significantly more complex than its single-echelon counterpart, aims to coordinate inventory decisions across multiple tiers such as factories, distributors, and retailers. This process involves optimizing inventory levels and replenishment policies at each tier to minimize total supply chain costs and mitigate systemic issues like the bullwhip effect.

ML approaches continue to offer solutions. For instance, Preil [53] developed a MCTS based algorithm for both offline and online models. Their method is designed to determine optimal inventory decisions under stochastic demand and lead times and also helps to alleviate the bullwhip effect. Zietsmann [68] introduced an integrated inventory management framework. This framework, built upon a computer-aided decision support system, cohesively handles product selection, demand forecasting, and procurement to optimize inventory under multi-objective constraints.

On the other hand, DL methods, particularly RL one, have become a cornerstone of modern multi-echelon inventory control, because they can handle more complex and dynamic cases. Zhang [52] proposed an improved SAC-AlphaLR algorithm to optimize replenishment policies in the supply chains of multi-tier fast-moving consumer goods. Addressing systemic challenges like the bullwhip effect is another critical application. Liu [63] utilized a HAPPO algorithm within a multi-agent deep RL framework. Their goal was to mitigate the bullwhip effect in distributed multi-echelon systems. Scalability is a major advantage offered by DL. Meisheri [25] developed a scalable deep RL approach that leverages a single-product meta-model to manage the inventory of hundreds of products in parallel. This method effectively handles multi-period constraints and demonstrates the potential of DL in

TABLE III. LITERATURE FOR INVENTORY CONTROL

Study	Citations	Type	Algorithms	Goals/Approach
[52]	34	DL	SAC-AlphaLR, DRL	A DRL-based ME-DRFO model is proposed for multi-echelon FMCG inventory optimization, whereby total costs are minimized under demand uncertainty and supply chain constraints.
[53]	94	ML	MCTS	An MCTS approach is proposed for multi-echelon inventory optimization to reduce total supply chain costs and mitigate the bullwhip effect.
[54]	14	Hybrid	GBDT, ANN	Inventory, ordering, and financing decisions are optimized using data-driven methods under capital constraints and demand volatility.
[55]	12	DL	DRL	High-stochasticity inventory optimization is addressed using a tailored DRL algorithm.
[56]	5	DL	PPO, DRL	A DRLOM is proposed for multi-echelon inventory control, by which system-wide costs under stochastic demand and lead times are minimized.
[57]	2	Others	System Dynamics	Replenishment is optimized using system dynamics, leading to a reduction in stockouts and lost sales.
[58]	11	DL	LSTM, PSO	Inventory errors and operational costs are minimized through Transformer-based forecasting and PSO.
[59]	17	DL/ML	DQN, LSTM, SVR	SME inventory is optimized through joint demand forecasting and replenishment policies.
[27]	28	DL	DQN	Multi-perishable inventory is optimized using deep RL, with lead times, lifecycles, and demand being considered.
[60]	76	DL	DRL, BNN, MAML	New-product inventory across multiple stores is optimized under demand uncertainty and short life cycles.
[61]	28	ML	IPSO, BPNN	Green supply chain inventory is optimized using an IPSO-BPNN algorithm to reduce costs and improve environmental efficiency.
[62]	17	ML	DTL	Deteriorating product inventory is optimized using ML forecasts and fuzzy variables to minimize cost and emissions.
[63]	17	DL	MADRL	Decentralized multi-echelon inventory optimization is addressed using multi-agent DRL with limited information sharing.
[64]	10	ML	RL	Online RL methods are proposed for inventory control under nonstationary demand, through which costs are saved and service levels are improved.
[65]	29	DL	DRL	Apparel inventory control is optimized using DRL, where demand volatility, service level, and sell-through rate are considered to minimize costs and improve efficiency.
[26]	126	DL	DRL	A DRL method is proposed for hospital drug inventory optimization, aimed at minimizing refilling costs under demand uncertainty and storage constraints.
[66]	25	DL	PPO, DRL	DRL is applied to two-echelon inventory control with stochastic seasonal demand and multiple warehouses, minimizing total costs.
[67]	45	ML	KNN, DT, RF, XGBoost	Hospital drug inventory is optimized using ML and stochastic optimization to reduce stock levels and emergency orders.
[22]	111	DL	DQN, TL	The problem of perishable inventory optimization under lead time and demand uncertainty is solved using DRLs with reward shaping.
[25]	41	DL	DQN	A DRL method is proposed for scalable multi-product inventory control with lead time constraints to optimize replenishment and reduce costs.
[24]	114	DL	DQN	The problem of inventory optimization for lost-sales and multi-echelon systems is solved using DRL, which minimizes costs under lead times and demand uncertainty.

large-scale supply chains. Beyond algorithmic development, researchers have also explored hybrid models. Xie [69] introduced the NFDIRM, which combines a RBF neural network with fuzzy logic. This fusion enhances both forecasting accuracy and decision flexibility in intelligent warehousing. Finally, providing a broader perspective, Boute [70] summarized key design choices in this domain and outlines promising directions for future research.

3) Inventory control in special scenarios: Inventory control in specialized contexts, such as for perishable goods and healthcare supplies, introduces distinct challenges due to strict temporal and safety constraints. RL has emerged as a powerful tool for addressing these complex problems.

For perishable goods, research has focused on dynamic policy learning to minimize spoilage and other losses. For instance, Selukar [27] developed a multi-product DQN model that dynamically learns optimal replenishment policies. Building on this foundation, De [22] improved the training stability of such models by incorporating reward shaping and transfer learning techniques. To better manage inherent uncertainties, Singh [62] adopted a fuzzy logic approach, which treats key parameters like deterioration and defect rates as fuzzy variables rather than fixed values.

Healthcare inventory management places a high premium on safety, and preventing stockouts is often a higher priority than minimizing costs. Addressing this critical balance, Abu [26] employed a deep RL framework for automated replenishment, which successfully reduced both operational costs and stockout incidents. Galli [67] integrated multiple ML algorithms with stochastic optimization, where the system leverages daily ward-level data to optimize inventory.

C. Inventory Classification

Inventory classification is a systematic methodology that categorizes inventory items based on criteria such as value, demand patterns, and criticality. This approach enables organizations to allocate resources efficiently by applying differentiated management policies to distinct categories. Table IV summarizes recent research on inventory classification.

1) Multi-criteria inventory classification: Although the traditional ABC classification method remains widely adopted, its reliance on a single metric, such as annual usage value, often neglects other critical operational factors [80]. To overcome this limitation, multi-criteria inventory classification (MCIC) has been introduced as a more robust approach. This paradigm integrates multiple indicators and thus provides a more detailed and effective basis for decision support.

One major research stream of MCIC involves clustering algorithms. For example, Keskin [72] combined an autoencoder with K-means, incorporating sustainability indicators such as carbon emissions alongside traditional measures like order fulfillment and inventory turnover. Similarly, Khanorkar [84] applied K-means to simultaneously optimize usage value, delivery time, and unit cost. Another important line of work involves multi-criteria decision-making (MCDM) techniques. Kaabi [77] integrated genetic algorithms with TOPSIS to systematically generate optimal weights for multiple objectives. Furthermore, to address inherent uncertainties, Yung

TABLE IV. LITERATURE FOR INVENTORY CLASSIFICATION

Study	Citations	Type	Algorithms	Goals/Approach
[71]	10	ML	K-means Clustering	The inventory classification problem for impulse purchase products is solved by grouping SKUs via unsupervised clustering (K-means).
[72]	1	ML	Autoencoder/K-means	The problem of inaccurate multi-criteria inventory classification in the white goods industry is solved by hybridizing autoencoders with clustering methods.
[73]	34	ML	SVM, MNB, DT, RF, GB, KNN	Inventory classification in blood bank supply chains is addressed using ML to automate data handling, with big data and security factors being considered.
[74]	2	DL	CapsNet	Automated consumer product identification and classification for inventory management is solved, with noise reduction, feature extraction, hyperparameter tuning, and ensemble learning being considered to maximize accuracy.
[75]	26	ML	SHAP	The lack of explainability in multi-criteria ABC inventory classification is addressed by proposing an XAI-based approach using SHAP values, optimizing both classification accuracy and interpretability.
[76]	8	ML	CNN	Inventory classification challenges in electrical warehouses are solved by blending RGB and depth CNN pipelines with synthetic and real data; classification accuracy is optimized to 95.23% while domain shifts are addressed.
[77]	15	ML	Genetic Algorithm	Multi-criteria inventory classification (ABC analysis) is addressed by integrating metaheuristics (AI) for criteria weighting and TOPSIS (MCDM) for scoring; total safety stock cost and service level are optimized.
[78]	19	DL	CNN	The problem of fabric type classification for inventory management is solved using YOLOv10.
[79]	10	ML	BPNN	An objective spare parts importance evaluation method is proposed to optimize inventory strategies and reduce subjectivity.
[80]	13	ML	SHAP/K-means	The lack of Pareto compliance and explainability in multi-criteria ABC inventory classification is solved by integrating semi-supervised clustering and XAI to provide interpretable results.
[81]	50	ML	Fuzzy Logic	A fuzzy ABC classification system for space mission components is proposed, which incorporates multi-attribute criteria to optimize purchasing decisions.
[82]	1	DL	CNN/Transformer	Fine-grained product recognition for inventory classification is addressed. High interclass similarity and large intraclass variation are tackled, and accuracy is optimized with self-supervised learning to reduce annotation dependency.
[83]	28	DL	CNN/Transformer	Multi-criteria spare parts classification problems are solved by transforming hierarchical criteria into visual graphs, and inventory strategies are optimized through a modified CNN.
[84]	34	ML	K-means	The problem of inaccurate inventory classification is solved by incorporating criteria like annual usage value, lead time, and unit cost with K-Means clustering.
[85]	30	ML	GA/DT	A cost-driven ML classification method is proposed, integrating demand patterns and replenishment policies to minimize inventory costs.

[81] employed fuzzy logic, which allowed them to manage variability in delivery times and supplier risks, thereby improving classification robustness. Khan [86] proposed a fuzzy set enhanced VIKORSort method that uses trapezoidal fuzzy membership functions to improve classification flexibility and decision-making efficiency. In addition, Medina [87] applied a genetic algorithm for ABC classification of raw material inventory in a jewelry company, and the method reduced the total inventory costs across multiple criteria and established differentiated inventory policies.

2) Enhanced classification interpretability: The 'black box' nature of learning algorithms presents a challenge in inventory classification. Although these models can achieve high accuracy, their opaque decision-making processes hinder trust and adoption. This is particularly critical in multi-criteria ABC analysis, where stakeholders must understand the reasoning behind classifications. Enhanced classification interpretability directly addresses this problem.

Research in this area primarily focuses on explainable artificial intelligence (XAI) and advanced visualization. The SHAP framework is an example of XAI. As demonstrated by Qaffas [75], its primary function is to quantify the contribution of each input feature to a model's final prediction. This allows for explanations at both the local level for individual items and the global level for the overall model. In their study [80], Qaffas further enhanced this approach by integrating semi-supervised clustering. This integration not only provides granular, multi-level explanations but also ensures that the classification results adhere to Pareto principles. Meanwhile, visualization techniques aim to make complex multi-criteria relationships intuitively understandable. For example, Yang [83] transformed hierarchical data structures into images and employed an improved CNN architecture to visualize decision boundaries and feature importance, and this method enhanced the transparency of the classification process. Moreover, Medina [88] proposed an integrated inventory management model that combines ABC-AHP classification, holt forecasting, and cycle counting. The model is designed to optimize inventory control and minimize stock discrepancies for distributors.

3) Classification in specified scenarios: Beyond the foundational principles of inventory classification, many sectors present unique operational requirements. In the healthcare sector, studies focus on managing life-critical items. For example, Maathavan [73] proposed a machine learning framework for blood banks that integrates parameters like expiration dates and storage protocols. This approach serves to minimize human error in high-stakes environments. In retail, the primary challenge stems from demand volatility. Garcia [71] addressed this with a customized K-means clustering method. The approach groups SKUs by demand and cost patterns to effectively manage unpredictable impulse-buy categories and optimize total relevant costs. In manufacturing and warehousing, object detection techniques are pivotal for real-time classification. For instance, the YOLOv10-based textile inventory system from Mao [78] reached 85.6% mAP. Similarly, to handle uncontrolled environments, Piratelo [76] used a hybrid of RGB and depth CNNs for warehouse classification. Their model proved highly robust, with an accuracy of 95.23% despite domain shifts.

D. Other AI Techniques in Inventory Management

AI in inventory management typically focuses on core tasks like demand forecasting, inventory control, and inventory classification. However, researchers are also applying AI to a wider range of research. This section explores several of these novel topics.

1) Automated simulation modeling: Traditional logistics simulation modeling heavily depends on close collaboration

between domain experts and specialized engineers, which creates barriers to its widespread adoption in inventory management. To address this issue, Jackson [89] developed a system based on an enhanced GPT-3 Codex model. This system utilizes advanced NLP and Transformer architectures to translate natural language descriptions into executable Python simulation code, which simplified the model development process.

2) 3D inventory positioning: The objective of 3D inventory positioning is to achieve precise automatic localization of inventory items in unstructured warehouses with flexible storage layouts. Traditional marker-based systems are often unsuitable for such dynamic environments due to their rigid infrastructure requirements. To overcome this limitation, Yoon [90] proposed a DL framework that processes drone-captured video from a single QR-coded camera. The framework integrates U-Net++ for image segmentation with epipolar geometry, resulting in a 17% improvement in coordinate estimation accuracy.

3) Robot navigation and task management: Efficient robot navigation and task management are crucial for automating modern warehouse operations. Research in this area addresses challenges ranging from physical pathfinding and environmental perception to high-level strategic decision-making.

One key challenge is maintaining mutual visibility and occlusion-free paths for cooperative autonomous systems, particularly for small UAVs with limited hardware. To address this problem, a visibility-aware framework was developed. Their system employs a cross-entropy method and a conditional variational autoencoder, which together significantly reduce occlusion and collision rates while maintaining real-time performance on embedded hardware. Beyond physical navigation, optimizing robot decision-making for tasks such as shelf reallocation is essential for adapting to dynamic inventory levels and traffic patterns. In [91], a DQN framework was proposed that integrates task assignment with shelf reallocation. This approach enables robots to optimize both task selection and storage repositioning concurrently, which resulted in a 17-23% increase in throughput and a 30% reduction in runtime compared to conventional methods.

4) Fruit quality monitoring: Food spoilage in micro-cold storage is a critical challenge for small-scale agriculture, where post-harvest losses can reach 35 to 40% due to inadequate quality monitoring. To mitigate such losses, Arun et al. [92] have proposed an AI-based system for real-time fruit freshness classification. Their approach utilizes the Inception architecture to analyze produce, and it demonstrated a classification accuracy of 99.78% under controlled laboratory conditions on apples, bananas, and oranges.

E. Application Examples

Beyond research, AI-driven solutions are being actively deployed to industries to solve problems. This section highlights specific applications of AI models across sectors such as energy, retail, manufacturing, healthcare, and agriculture. These examples showcase how AI-driven approaches address domain-specific challenges.

1) Energy sector: The energy sector faces challenges in material scheduling and managing demand uncertainty. To optimize material scheduling, Hu [93] proposed an NSGA-II algorithm that integrates non-dominated sorting with rough vector feature decomposition. This method reduced convergence iterations from 460 to 65 and improved dispatch reliability. In addressing demand uncertainty, Thoummala [94] developed a two-stage DL framework. The model employs an LSTM network for forecasting to determine optimal inventory levels, resulting in inventory cost reductions ranging from 4.16% to 12.10%.

2) Retail sector: In the retail sector, key challenges include multi-echelon inventory coordination and the implementation of dynamic pricing strategies. To address coordination, Kaynov [95] proposed a DRL solution based on a PPO framework. This approach reduced network complexity and alleviated shortage bias, leading to cost reductions of 1 to 3% in lost-sales scenarios and 10 to 20% in backordering scenarios. For dynamic pricing, Liang [37] introduced a system based on DQN, modeling sales cycles as a Markov decision process. By decoupling the value and action-advantage functions, the algorithm achieved 91.58% of the optimal theoretical performance.

3) Manufacturing sector: Manufacturing supply chains are often hindered by demand information asymmetry, high inventory costs, and response delays. To mitigate the bullwhip effect, Sardar [96] developed a system that integrates an LSTM model with RFID tracking and a consignment strategy. This system enabled precise inventory control, and brings profit increase of 43%. In specialized fields such as pharmaceuticals, where historical data is limited, Zhu [97] leveraged multi-product data and grouping strategies to capture complex demand patterns more effectively.

4) Healthcare sector: Inventory management in healthcare must address the challenges of perishable resources, such as blood products and pharmaceuticals, where demand uncertainty and static policies can cause costly expirations or critical shortages. To optimize blood inventory, Mohamadi [98] developed a framework based on an A2C RL algorithm. This framework enabled dynamic distribution policies that achieved a zero-expiration rate while maintaining an 85% service level. To mitigate drug shortages, Abu [26] proposed a DQN-based replenishment model that formulates the problem as a Markov decision process. The model balances routine inventory with emergency demand, which reduced the shortage rate to 2.21% and the overall costs by 12.31%. Additionally, Adedugba [99] developed a dual-scenario deterministic inventory model for a pharmaceutical manufacturer in Lagos, Nigeria, using mathematical optimization to improve inventory hierarchies, lead times, and shortage management.

5) Agriculture sector: Agricultural supply chains face severe challenges, including perishable product losses of up to 35–40% and significant seasonal imbalances between supply and demand. For freshness monitoring, Arun [92] developed an Inception v3-based algorithm to process sensor-captured images, achieving 99% classification accuracy. This enabled more precise harvest timing and contributed to price stabilization. To optimize seasonal inventory, Wang [100] proposed a DQN framework integrated with real-time IoT data. The model incorporates reward shaping and cosine functions to effectively capture seasonal demand patterns.

V. Discussion

A. RQ1: What is the Current State of Research on AI Technologies in Inventory Management?

To address this question, in Section III, this study analyzed publication trends, source distributions, and high-impact studies using a predefined set of subject terms, keywords, and application domains.

Publication trends indicate a continuous increase in AI-driven inventory optimization research from 2020 to 2025. As shown in Fig. 2, a substantial surge in publications occurred in 2022, followed by stable growth over the next three years. This trajectory reflects the growing recognition of AI's potential to enhance inventory performance, while rapid technological advancements have further promoted practical adoption.

An analysis of keywords reveals the field's main research areas. The most prominent topics are demand forecasting and inventory management, which together account for nearly 28% of the research. This high percentage shows the foundational role of forecast accuracy in effective inventory optimization. Next, inventory control and optimization make up another 24%, which indicates that translating forecasts into actionable strategies is still a central research challenge. Other major research topics include dynamic pricing, replenishment, and supply chain resilience, with each representing about 12% of the literature. This focus reflects significant scholarly attention on profit maximization and risk mitigation. In contrast, topics like multi-criteria classification, warehouse automation, and new technologies such as AI and RL currently represent a smaller but growing area of research.

Several studies have demonstrated significant influence in this domain (see Table I). For instance, [22] has garnered substantial attention with 126 citations, followed by [23] (115 citations) and [24] (114 citations). These highly cited works underscore the importance of forecasting and advanced optimization strategies. Overall, the top ten cited articles primarily focus on optimization algorithms, DL, and explainable AI, reflecting sustained efforts to leverage advanced technologies for improved inventory management efficiency and decision support.

B. RQ2: What are the Primary Topics in Inventory Management that AI is Applied to Solve?

Although AI has been widely applied to a wide range of topics, this study finds that its most significant impact is on three core challenges: demand forecasting, inventory control, and inventory classification. These domains form the operational foundation of modern inventory strategies, where AI demonstrates the greatest value in overcoming the limitations of traditional methods.

1) Demand forecasting: Demand forecasting aims to predict future demand accurately by modeling complex patterns, such as nonlinear trends and temporal dependencies, from historical data and diverse features. A key aspect of modern forecasting is not only to provide point estimates but also to quantify uncertainty through techniques like interval prediction [48], [46], [42], allowing for robust adaptation across various industries and contexts .

However, achieving this goal in practice faces fundamental challenges. A primary issue is that these models often fail to capture the complex nonlinearities and multivariate dependencies that drive demand [33], [43]. Second, their single point estimates neglect the quantification of forecast uncertainty, which is a critical input for robust inventory and risk management [46]. Finally, their effectiveness is frequently undermined by the practical realities of sparse, incomplete, or low-quality data.

To address these forecasting challenges, recent research has shifted towards AI-driven methods capable of capturing complex demand patterns. Deep learning models, particularly LSTM networks and their variants, have proven effective at modeling long-term temporal dependencies. For instance, Hasni [33] demonstrated that a hybrid LSTM model could reduce prediction errors for intermittent medical supplies by 74.14%.

Further advancements include the integration of attention mechanisms and sophisticated optimization algorithms. A notable example is the BO-CNN-LSTM model, which employs Bayesian optimization to automatically fine-tune the model architecture, simultaneously enhancing forecasting precision and inventory cost control [45]. Beyond improving accuracy, researchers have also tackled data scarcity. Generative AI and transfer learning are now being explored to augment sparse datasets [42], which is crucial for ensuring robust predictions for new products or in settings with limited historical data.

2) Inventory control: Inventory control dynamically determines optimal ordering and stocking policies to balance the trade-off between minimizing costs and maximizing service levels, while adapting to complex constraints and demand uncertainties across the supply chain [53], [52], [27], [22].

However, traditional inventory control methods are often ill-equipped for modern complexities. They struggle with diverse, non-stationary demand patterns, and a structural disconnect between forecasting and control allows prediction errors to propagate directly into costly stock imbalances [53], [52], [59]. These demand-side issues are further amplified by supply chain dynamics. Within the supply chain, information asymmetries can fuel instabilities such as the bullwhip effect. Furthermore, a complex web of real-world constraints often renders simple optimization models ineffective.

To address these challenges, DRL has emerged as a powerful paradigm for developing dynamic ordering policies that sense and adapt to non-stationary demand in real time. These policies have shown superior performance over static models [52], [53]. A key architectural innovation is the integrated forecast-and-optimize framework. In this framework, deep learning models first capture complex demand patterns. Their outputs then directly inform optimization algorithms, which prevents the propagation of prediction errors [58], [59]. This principle of integration extends to multi-echelon systems. Here, multi-agent RL allows for the development of collaborative policies that coordinate decisions across all supply chain nodes. This approach effectively dampens the bullwhip effect and aligns local actions with global system efficiency.

3) Inventory classification: Inventory classification is a critical process that segments items based on multiple criteria to align management strategies with their strategic importance.

However, conventional methods face significant practical limitations. First, they are constrained by an over-reliance on single or subjective criteria. Classic ABC analysis, for instance, uses a single metric and neglects critical operational factors from supplier risk to sustainability [72]. Even multi-criteria approaches often depend on subjective expert weights, which undermines objectivity and auditability [77]. Second, their effectiveness is compromised by opaque "black-box" logic. As models grow in complexity, their internal decision-making becomes unintelligible to managers, eroding trust and hindering practical adoption [80]. Third, they exhibit poor adaptability to real-world data, and are difficult to manage the uncertainty, noise, or sparsity common in modern inventories.

To overcome these challenges, AI-driven methods offer targeted advancements. To counter subjective bias, AI employs data-driven multi-objective optimization. Techniques like K-means clustering and genetic algorithm-enhanced MCDM autonomously learn the importance of diverse metrics, establishing an objective, evidence-based framework [72], [77]. To resolve the 'black-box' problem, AI incorporates eXplainable AI (XAI) techniques. Frameworks such as SHAP deconstruct model decisions, and provides transparent, quantifiable justifications for item categorization that enhance trust and stakeholder buy-in [80]. Finally, to tackle poor data adaptability, AI leverages advanced data processing. Fuzzy logic [81], [72] is used to effectively integrate mixed criteria, while feature enhancement tools like autoencoders ensure model robustness against noise and data sparsity.

C. RQ3: What are the Key AI Techniques and Algorithms Applied in Inventory Management, and What are their Respective Advantages and Disadvantages?

According to the review, AI techniques in inventory management can be categorized into traditional ML, DL, RL, and hybrid algorithms.

1) ML methods: Algorithms such as KNN, DT, RF, XGBoost, LightGBM, SVR, ANN have been widely adopted in inventory management. These methods are primarily applied to demand forecasting and can handle volatile demand [23], [29], [36]. They also support basic inventory decisions, such as optimizing replenishment quantities for profit maximization [61], [67]. The main advantages of ML algorithms lie in their low computational costs and fast inference speeds. In addition, they provide good interpretability and can clearly identify the contribution of each feature. However, their ability to capture long-term dependencies is limited. For example, SVR often performs poorly in modeling multi-period demand for short-shelf-life products [47]. In addition, ML models frequently require substantial manual intervention [34], including parameter tuning and feature engineering, and their generalization ability is constrained.

2) DL methods: Models including LSTM, CNN, Transformers, and AEs are increasingly applied to complex inventory management problems. They are primarily used for advanced demand forecasting, capable of handling non-stationary, intermittent, and multi-factor-driven demand patterns, such as forecasting procurement needs for perishable goods [35], and multi-category inventory requirements influenced by factors like weather and promotions [45]. DL models excel at high-dimensional feature extraction, automatically capturing temporal dependencies and multi-feature interactions. For instance, MMCW-BiLSTM [43] captures bidirectional temporal relationships, while Transformers identify long-range nonlinear dependencies.

The main advantage of DL lies in its superior predictive accuracy, which significantly outperforms traditional statistical methods. For example, in [58], a Transformer model reduced the MAE of the predicted values from 15.8 to 8.2 and RMSE from 22.3 to 11.5. DL models also demonstrate strong adaptability to complex demand patterns. In [43], BiLSTM can process nonlinear time-series data with MAE as low as 1.75. Their generalization ability is also notable, with MMCW-BiLSTM and BO-CNN-LSTM achieving high accuracy across diverse datasets and scenarios [45].

However, compared with ML methods, DL models have high computational costs, and higher dependence on large-scale, high-quality datasets. Moreover, the 'black-box' nature of these models makes decision logic difficult to trace, which restrict their applicability in regulated fields such as healthcare.

3) RL methods: The RL algorithms, such as PPO, A2C, DDPG, and DQN, have emerged as powerful tools for dynamic inventory control. RL is particularly well-suited for multi-echelon supply chain management, perishable and multi-category inventory control, and joint inventory-financing decision-making under constraints [26], [55]. A key application is mitigating the bullwhip effect by promoting coordination and information sharing among supply chain nodes. For instance, the HAPPO [44] model can reduce demand amplification by 20%.

The core advantage of RL is its strong dynamic adaptability. It provides responses to stochastic demand, lead-time variability, and changing inventory states to achieve significant cost optimization. For example, in [53], the supplier order variability is reduced by 21%. In addition, its high-dimensional handling capability enables scalability for large-scale SKU management, with inference speeds one to two orders of magnitude faster than linear programming [53]. However, RL comes with substantial challenges: training costs are high, with models such as PPO and A2C requiring millions of iterations and several hours to converge [27] [56]. Additionally, multi-agent coordination in multi-echelon supply chains remains a challenge, as MARL models often exhibit slow convergence and struggle to balance the objectives of individual nodes with the global optimum.

4) Hybrid algorithms: Hybrid algorithms combine the strengths of multiple AI techniques to create more powerful and flexible solutions. Examples include LSTM-PSO [60], IPSO-BPNN [61], and joint 'prediction + inventory' frameworks such as W-LSTM + Q-learning [59]. These models are designed for 'forecast-then-optimize' workflows, bridging the gap between demand prediction and inventory decision-making to enhance overall efficiency. Phatangare proposed a machine learning based food inventory management system, leveraging linear regression, random forest, and LSTM networks [101] to optimize inventory levels through demand forecasting.

The primary advantage of hybrid models is their synergy, which combines the strengths of different techniques

to achieve superior performance. For instance, the LSTM-PSO model integrates the forecasting accuracy of LSTM with the optimization capability of PSO. This integration reduces fulfillment time by 25% and improves both model accuracy and robustness [58]. Another key benefit is their adaptability to complex scenarios. However, these advantages come at the cost of structural complexity and high computational demands, a primary disadvantage evident in the increased training times required for AE-Clustering hybrids.

D. RQ4: What Challenges does AI Face in Inventory Management, and What Directions can be Explored in Future Research?

The integration of artificial intelligence into inventory management also faces challenges spanning data dependency, model architecture, and operational viability.

First, the performance of AI models is critically constrained by data dependency. Most algorithms presuppose the availability of large, high-quality datasets and are highly sensitive to common data imperfections such as noise, outliers, and incompleteness [80]. A more profound limitation is their frequent inability to adapt to the complex data landscapes characteristic of real-world inventory systems: the heterogeneous, imbalanced, and non-stationary data associated with scarce, high-value components or sudden market disruptions [24][29]. Second, significant challenges are inherent to the architecture and functionality of current AI models. The inherent 'black box' nature of DL models impedes their adoption in high-stakes environments where decision transparency and interpretability are paramount [75]. Furthermore, many models exhibit narrow functional scopes. For instance, a DRL agent optimized for policy control may fail to manage volatile demand without an integrated forecasting component [60]. Finally, operational costs present a barrier to widespread adoption. For example, baseline models such as A2C [27] and DDPG [84] can require millions of simulation steps to achieve convergence.

To overcome these challenges, future research may focus on several key directions:

- A primary avenue is the development of data-efficient and robust learning paradigms. This includes leveraging frontier technologies like the Internet of Things (IoT) and Computer Vision to capture high-fidelity, real-time data, thus mitigating issues of scarcity and quality. Furthermore, the application of Generative AI and LLMs holds immense potential for creating high-fidelity simulation environments and synthetic training data. Such approaches can augment sparse datasets, balance class distributions, and simulate black swan events, thereby enhancing model resilience and generalization in dynamic conditions.
- Another critical research frontier is the pursuit of explainable, integrated, and automated modeling frameworks. Incorporating explainable AI techniques is crucial for demystifying model predictions, allowing managers to understand the rationale behind AI-driven decisions and fostering greater trust.
- Finally, to tackle the prohibitive operational costs, future research must move towards more efficient and lightweight models. A key direction is to develop algorithms that learn faster from less data, such as advancing offline reinforcement learning to leverage existing historical records instead of relying on extensive live simulations.

VI. Conclusion

This systematic review consolidates the current landscape of AI in inventory management by analyzing 87 articles published between 2021 and 2025. Our main finding is that AI applications primarily focus on three core areas. The first is demand forecasting, where deep learning is effective at capturing complex patterns. The second is inventory control, where reinforcement learning is increasingly used to handle dynamic and uncertain environments. The third is inventory classification, where explainable AI makes traditional methods more transparent. We also identified a consistent trade-off among different AI approaches. Machine learning is computationally efficient, while deep learning offers high accuracy but requires large amounts of data. Reinforcement learning allows for real-time adaptation but demands significant training resources. These results highlight both the great potential and the practical challenges of using AI for inventory management.

The novelty of this study is its focused and timely analysis of literature published after 2020. Unlike broader reviews, our work provides a detailed map of specific AI techniques to fundamental inventory problems. This mapping clarifies the distinct advantages and trade-offs of each technique. A key contribution of our study is the identification of a clear agenda for future research aimed at overcoming current limitations. By suggesting a focus on data-efficient learning, integrated prediction and decision models, and lightweight algorithms, this study guides future academic research and industry applications.

Despite these contributions, this study has several limitations. Its timeframe may have excluded foundational research published before 2021. In addition, our keyword-based search strategy might have missed relevant studies that used different terminology. These weaknesses indicate that our conclusions offer a comprehensive but not exhaustive overview of the field, which can be further strengthened in future studies.

References

[1] Mugdha, D. Sharma, and D. Jose, "Transforming inventory management through machine learning based demand forecasting for a circular economy," in *2024 International Conference on Communication, Control, and Intelligent Systems (CCIS)*, 2024, pp. 1–6.

[2] X. Kong, J. Yang, W. Chen, and J. Pan, "Joint multimission selective maintenance and inventory optimization for multicomponent systems considering stochastic dependency," *IEEE Transactions on Reliability*, vol. 73, no. 4, pp. 1967–1981, 2024.

[3] F. Nurprihatin, Y. T. Prasetyo, R. Chandra, I. T. B. Widiwati, T. E. Lestari, and S. S. Islam, "Strategic inventory control using economic order quantity and abc classification to enhance operational excellence," in *2025 4th International Conference on Computational Modelling, Simulation and Optimization (ICCMSO)*, 2025, pp. 40–46.

[4] V. O. Adebajo-Akinpelu, M. O. Onibonoje, S. L. Gbadamosi, A. O. Femi-Onibonoje, A. S. Yusuff, and T. M. Azeez, "Enhancing inventory and payment management: A review of current technologies and future direction," in *2024 IEEE 5th International Conference on Electro-Computing Technologies for Humanity (NIGERCON)*, 2024, pp. 1–4.

[5] S. Impram, S. V. Nese, and B. Oral, "Challenges of renewable energy penetration on power system flexibility: A survey," *Energy strategy reviews*, vol. 31, p. 100539, 2020.

[6] Q. Tan, X. Li, and Y. Liang, "Risks, challenges and strategies of power systems against the background of carbon neutrality," *Clean Energy*, vol. 7, no. 4, pp. 767–782, 2023.

[7] C.-y. Liu and Y.-f. Sun, "Modeling process integrated quality management system in manufacturing enterprises," in *2009 Second International Conference on Future Information Technology and Management Engineering*, 2009, pp. 77–80.

[8] S. Ruidas, M. Rahaman Seikh, and P. Kumar Nayak, *Application of particle swarm optimization technique in an interval-valued EPQ model*, 2022, pp. 51–78.

[9] M. Khaleel, A. Jebrel, and D. M. Shwehdy, "Artificial intelligence in computer science: Https://doi. org/10.5281/zenodo. 10937515," *Int. J. Electr. Eng. and Sustain.*, pp. 01–21, 2024.

[10] M. K. A. Ismaeil, "The role and impact of artificial intelligence on supply chain management: efficiency, challenges, and strategic implementation," *Journal of Ecohumanism*, vol. 3, no. 4, pp. 89–106, 2024.

[11] M. Pournader, H. Ghaderi, A. Hassanzadegan, and B. Fahimnia, "Artificial intelligence applications in supply chain management," *International Journal of Production Economics*, vol. 241, p. 108250, 2021.

[12] P. Sharma, A. Gunasekaran, and G. Subramanian, "Enhancing supply chain: exploring and exploiting ai capabilities," *Journal of Computer Information Systems*, pp. 1–15, 2024.

[13] G. Manoharan, A. Sharma, V. D. Vani, V. H. Raj, R. Jain, and G. Nijhawan, "Predictive analytics for inventory management in e-commerce using machine learning algorithms," in *2024 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*. IEEE, 2024, pp. 1–5.

[14] T. Yavuz and O. Kaya, "Deep reinforcement learning algorithms for dynamic pricing and inventory management of perishable products," *Applied Soft Computing*, vol. 163, p. 111864, 2024.

[15] J. Zhu, S. A. Bazaz, S. Dutta, B. Anuraag, I. Haider, and S. Bandopadhyay, "Talk to your data: Enhancing business intelligence and inventory management with llm-driven semantic parsing and text-to-sql for database querying," in *2023 4th International Conference on Data Analytics for Business and Industry (ICDABI)*, 2023, pp. 321–325.

[16] K. Wang, C. Long, D. J. Ong, J. Zhang, and X.-M. Yuan, "Single-site perishable inventory management under uncertainties: A deep reinforcement learning approach," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 10, pp. 10 807–10 813, 2023.

[17] W. Villegas-Ch, A. M. Navarro, and S. Sanchez-Viteri, "Optimization of inventory management through computer vision and machine learning technologies," *Intelligent Systems with Applications*, vol. 24, p. 200438, 2024.

[18] L. Chen, S. Zhang, N. Wu, Y. Qiao, Z. Zhong, and T. Chen, "Optimization of inventory space in smart factory for integrated periodic production and delivery scheduling," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 6, pp. 3488–3511, 2023.

[19] P. Verma, "Transforming supply chains through ai: Demand forecasting, inventory management, and dynamic optimization," *Integrated Journal of Science and Technology*, vol. 1, no. 3, 2024.

[20] P. Gupta and A. Sachan, "Resource planning and inventory management in supply chain," in *Supply Chain Management*. CRC Press, 2024, pp. 58–79.

[21] Ö. Albayrak Ünal, B. Erkayman, and B. Usanmaz, "Applications of artificial intelligence in inventory management: A systematic review of the literature," *Archives of computational methods in engineering*, vol. 30, no. 4, pp. 2605–2625, 2023.

[22] B. J. De Moor, J. Gijsbrechts, and R. N. Boute, "Reward shaping to improve the performance of deep reinforcement learning in perishable inventory management," *European Journal of Operational Research*, vol. 301, no. 2, pp. 535–545, 2022.

[23] J. N. Gonçalves, P. Cortez, M. S. Carvalho, and N. M. Frazao, "A multivariate approach for multi-step demand forecasting in assembly industries: Empirical evidence from an automotive supply chain," *Decision Support Systems*, vol. 142, p. 113452, 2021.

[24] Q. Wang, Y. Peng, and Y. Yang, "Solving inventory management problems through deep reinforcement learning," *Journal of Systems Science and Systems Engineering*, vol. 31, no. 6, pp. 677–689, 2022.

[25] H. Meisheri, N. N. Sultana, M. Baranwal, V. Baniwal, S. Nath, S. Verma, B. Ravindran, and H. Khadilkar, "Scalable multi-product inventory control with lead time constraints using reinforcement learning," *Neural Computing and Applications*, vol. 34, no. 3, pp. 1735–1757, 2022.

[26] T. Abu Zwaida, C. Pham, and Y. Beauregard, "Optimization of inventory management to prevent drug shortages in the hospital supply chain," *Applied Sciences*, vol. 11, no. 6, p. 2726, 2021.

[27] M. Selukar, P. Jain, and T. Kumar, "Inventory control of multiple perishable goods using deep reinforcement learning for sustainable environment," *Sustainable Energy Technologies and Assessments*, vol. 52, p. 102038, 2022.

[28] Y. M. Tang, K. Y. Chau, Y.-y. Lau, and Z. Zheng, "Data-intensive inventory forecasting with artificial intelligence models for cross-border e-commerce service automation," *Applied Sciences*, vol. 13, no. 5, p. 3051, 2023.

[29] M. A. Jahin, M. S. H. Shovon, M. S. Islam, J. Shin, M. F. Mridha, and Y. Okuyama, "Qamplifynet: pushing the boundaries of supply chain backorder prediction using interpretable hybrid quantum-classical neural network," *Scientific reports*, vol. 13, no. 1, p. 18246, 2023.

[30] M. Shajalal, A. Boden, and G. Stevens, "Explainable product backorder prediction exploiting cnn: Introducing explainable models in businesses," *Electronic Markets*, vol. 32, no. 4, pp. 2107–2122, 2022.

[31] J. Xu and L. Bo, "Optimizing supply chain resilience using advanced analytics and computational intelligence techniques," *IEEE Access*, 2024.

[32] C. Deng and Y. Liu, "A deep learning-based inventory management and demand prediction optimization method for anomaly detection," *Wireless communications and mobile computing*, vol. 2021, no. 1, p. 9969357, 2021.

[33] M. Hasni, M. Z. Babai, and B. Rostami-Tabar, "A hybrid lstm method for forecasting demands of medical items in humanitarian operations," *International Journal of Production Research*, vol. 62, no. 17, pp. 6046–6063, 2024.

[34] L. A. C. Ahakonye, A. Zainudin, M. J. A. Shanto, J.-M. Lee, D.-S. Kim, and T. Jun, "A multi-mlp prediction for inventory management in manufacturing execution system," *Internet of Things*, vol. 26, p. 101156, 2024.

[35] C.-Y. Kao and H.-E. Chueh, "Deep learning based purchase forecasting for food producer-retailer team merchandising," *Scientific Programming*, vol. 2022, no. 1, p. 2857850, 2022.

[36] A. Aktepe, E. Yanık, and S. Ersöz, "Demand forecasting application with regression and artificial intelligence methods in a construction machinery company," *Journal of Intelligent Manufacturing*, vol. 32, no. 6, pp. 1587–1604, 2021.

[37] Y. Liang, Y. Hu, D. Luo, Q. Zhu, Q. Chen, and C. Wang, "Distributed dynamic pricing strategy based on deep reinforcement learning approach in a presale mechanism," *Sustainability*, vol. 15, no. 13, p. 10480, 2023.

[38] D. Li, F. Gu, X. Li, R. Du, D. Chen, and A. Madden, "Dynamic sales prediction with auto-learning and elastic-adjustment mechanism for inventory optimization," *Information Systems*, vol. 119, p. 102259, 2023.

[39] S. K. Jauhar, S. Harinath, V. Krishnaswamy, and S. K. Paul, "Explainable artificial intelligence to improve the resilience of perishable product supply chains by leveraging customer characteristics," *Annals of Operations Research*, pp. 1–40, 2024.

[40] T. Sathish, D. SaiKumar, S. Patil, R. Saravanan, J. Giri, and A. A. Aly, "Exponential smoothing method against the gradient boosting machine learning algorithm-based model for materials forecasting to minimize inventory," *AIP Advances*, vol. 14, no. 6, 2024.

[41] K. Lutoslawski, M. Hernes, J. Radomska, M. Hajdas, E. Walaszczyk, and A. Kozina, "Food demand prediction using the nonlinear autoregressive exogenous neural network," *IEEE Access*, vol. 9, pp. 146 123–146 136, 2021.

[42] M. Kim, J. Lee, C. Lee, and J. Jeong, "Framework of 2d kde and lstm-based forecasting for cost-effective inventory management in smart manufacturing," *Applied sciences*, vol. 12, no. 5, p. 2380, 2022.

[43] S. Abualuroug, A. Alzubi, and K. Iyiola, "Inventory prediction using a modified multi-dimensional collaborative wrapped bi-directional long short-term memory model," *Applied Sciences*, vol. 14, no. 13, p. 5817, 2024.

[44] Y. Yang, M. Wang, J. Wang, P. Li, and M. Zhou, "Multi-agent deep reinforcement learning for integrated demand forecasting and inventory optimization in sensor-enabled retail supply chains," *Sensors (Basel, Switzerland)*, vol. 25, no. 8, p. 2428, 2025.

[45] R. Liu and V. Vakharia, "Optimizing supply chain management through bo-cnn-lstm for demand forecasting and inventory management," *Journal of Organizational and End User Computing (JOEUC)*, vol. 36, no. 1, pp. 1–25, 2024.

[46] H. Cui, H. Guo, J. Wang, and Y. Wang, "Point and interval forecasting for wine prices: an approach based on artificial intelligence," *International Journal of Contemporary Hospitality Management*, vol. 36, no. 8, pp. 2752–2773, 2024.

[47] S. K. Panda and S. N. Mohanty, "Time series forecasting and modeling of food demand supply chain based on regressors analysis," *Ieee Access*, vol. 11, pp. 42 679–42 700, 2023.

[48] W. A. Zogaan, N. Ajabnoor, and A. A. Salamai, "Leveraging deep learning for risk prediction and resilience in supply chains: insights from critical industries," *Journal of Big Data*, vol. 12, no. 1, p. 94, 2025.

[49] A. Manos, D. E. Filippidou, N. Pavlidis, G. Karanasios, G. Vachtanidis, I. Mallidis, O. Tsolakis, and G. Ayfantopoulou, "Dynamic demand forecasting with ai: Maximizing supply chain value," in *2024 International Conference on Engineering and Emerging Technologies (ICEET)*, 2024, pp. 1–4.

[50] M. Seyedan, F. Mafakheri, and C. Wang, "Order-up-to-level inventory optimization model using time-series demand forecasting with ensemble deep learning," *Supply Chain Analytics*, vol. 3, p. 100024, 2023.

[51] T. Jin, "Attention-based temporal convolutional networks and reinforcement learning for supply chain delay prediction and inventory optimization," in *2024 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*. IEEE, 2024, pp. 1527–1531.

[52] Y. Zhang, L. He, and J. Zheng, "A deep reinforcement learning-based dynamic replenishment approach for multi-echelon inventory considering cost optimization," *Electronics*, vol. 14, no. 1, p. 66, 2024.

[53] D. Preil and M. Krapp, "Artificial intelligence-based inventory management: a monte carlo tree search approach," *Annals of Operations Research*, vol. 308, no. 1, pp. 415–439, 2022.

[54] B. Yang, X. Xu, Y. Gong, and Y. Rekik, "Data-driven optimization models for inventory and financing decisions in online retailing platforms," *Annals of Operations Research*, vol. 339, no. 1, pp. 741–764, 2024.

[55] T. Temizöz, C. Imdahl, R. Dijkman, D. Lamghari-Idrissi, and W. van Jaarsveld, "Deep controlled learning for inventory control," *European Journal of Operational Research*, vol. 324, no. 1, pp. 104–117, 2025.

[56] D. S. Kurian, V. M. Pillai, A. Raut, and J. Gautham, "Deep reinforcement learning-based ordering mechanism for performance optimization in multi-echelon supply chains," *Applied Stochastic Models in Business and Industry*, vol. 40, no. 5, pp. 1433–1454, 2024.

[57] A. Rohilla, T. Kundu, R. Kapoor, and J.-B. Sheu, "Enhancing retail inventory replenishment amid product life cycle shifts: A system dynamics approach," *IEEE Transactions on Engineering Management*, 2025.

[58] J. Xu and L. Bo, "Enhancing supply chain efficiency resilience using predictive analytics and computational intelligence techniques," *IEEE Access*, 2024.

[59] H. J. Wahedi, M. Heltoft, G. J. Christophersen, T. Severinsen, S. Saha, and I. E. Nielsen, "Forecasting and inventory planning: An empirical investigation of classical and machine learning approaches for svanehøj's future software consolidation," *Applied Sciences*, vol. 13, no. 15, p. 8581, 2023.

[60] T. Demizu, Y. Fukazawa, and H. Morita, "Inventory management of new products in retailers using model-based deep reinforcement learning," *Expert Systems with Applications*, vol. 229, p. 120256, 2023.

[61] Y. Guan, Y. Huang, and H. Qin, "Inventory management optimization of green supply chain using ipso-bpnn algorithm under the artificial intelligence," *Wireless Communications and Mobile Computing*, vol. 2022, no. 1, p. 8428964, 2022.

[62] R. Singh and V. K. Mishra, "Machine learning based fuzzy inventory model for imperfect deteriorating products with demand forecast and partial backlogging under green investment technology," *Journal of the Operational Research Society*, vol. 75, no. 7, pp. 1223–1238, 2024.

[63] X. Liu, M. Hu, Y. Peng, and Y. Yang, "Multi-agent deep reinforcement learning for multi-echelon inventory management," *Production and Operations Management*, vol. 34, no. 7, pp. 1836–1856, 2025.

[64] D. Y. Mo, Y. P. Tsang, Y. Wang, and W. Xu, "Online reinforcement learning-based inventory control for intelligent e-fulfilment dealing with nonstationary demand," *Enterprise Information Systems*, vol. 18, no. 2, p. 2284427, 2024.

[65] J. W. Chong, W. Kim, and J. Hong, "Optimization of apparel supply chain using deep reinforcement learning," *IEEE Access*, vol. 10, pp. 100 367–100 375, 2022.

[66] F. Stranieri, F. Stella, and C. Kouki, "Performance of deep reinforcement learning algorithms in two-echelon inventory control systems," *International Journal of Production Research*, vol. 62, no. 17, pp. 6211–6226, 2024.

[67] L. Galli, T. Levato, F. Schoen, and L. Tigli, "Prescriptive analytics for inventory management in health care," *Journal of the Operational Research Society*, vol. 72, no. 10, pp. 2211–2224, 2021.

[68] H. J. Zietsman and J. H. van Vuuren, "A generic decision support framework for inventory procurement planning in distribution centres," *Computers & Industrial Engineering*, vol. 177, p. 109057, 2023.

[69] C. Xie and C. Xie, "Inventory control strategy based on neural network and fuzzy algorithm in intelligent warehousing system," *Discover Artificial Intelligence*, vol. 5, no. 1, p. 159, 2025.

[70] R. N. Boute, J. Gijsbrechts, W. Van Jaarsveld, and N. Vanvuchelen, "Deep reinforcement learning for inventory control: A roadmap," *European journal of operational research*, vol. 298, no. 2, pp. 401–412, 2022.

[71] D. García-Barrios, "A machine learning based method for managing multiple impulse purchase products: an inventory management approach," *Journal of Engineering Science and Technology Review*, 2021.

[72] S. Keskin and A. Taskin, "A novel autoencoder-integrated clustering methodology for inventory classification: A real case study for white goods industry." *Sustainability (2071-1050)*, vol. 16, no. 21, 2024.

[73] K. S. K. Maathavan and S. Venkatraman, "A secure encrypted classified electronic healthcare data for public cloud environment." *Intelligent Automation & Soft Computing*, vol. 32, no. 2, 2022.

[74] R. Alabdan, H. K. Alkahtani, K. M. Othman, R. Alroobaea, S. Algarni, A. A. Alharbi, and K. Shankar, "Advanced imaging technologies with ensemble learning for consumer products identification and classification," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 4, pp. 7174–7182, 2024.

[75] A. A. Qaffas, M.-A. Ben HajKacem, C.-E. Ben Ncir, and O. Nasraoui, "An explainable artificial intelligence approach for multi-criteria abc item classification," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 18, no. 2, pp. 848–866, 2023.

[76] P. H. M. Piratelo, R. N. de Azeredo, E. M. Yamao, J. F. Bianchi Filho, G. Maidl, F. S. M. Lisboa, L. P. de Jesus, R. d. A. Penteado Neto, L. d. S. Coelho, and G. V. Leandro, "Blending colored and depth cnn pipelines in an ensemble learning classification approach for warehouse application using synthetic and real data," *Machines*, vol. 10, no. 1, p. 28, 2021.

[77] H. Kaabi, "Comparative analysis of multicriteria inventory classification models for abc analysis," *International Journal of Information Technology & Decision Making*, vol. 21, no. 05, pp. 1617–1646, 2022.

[78] M. Mao, A. Lee, and M. Hong, "Efficient fabric classification and object detection using yolov10." *Electronics (2079-9292)*, vol. 13, no. 19, 2024.

[79] S. Zhang, X. Qin, S. Hu, Q. Zhang, B. Dong, and J. Zhao, "Importance degree evaluation of spare parts based on clustering algorithm and back-propagation neural network," *Mathematical Problems in Engineering*, vol. 2020, no. 1, p. 6161825, 2020.

[80] A. A. Qaffas, M. A. B. Hajkacem, C.-E. B. Ncir, and O. Nasraoui, "Interpretable multi-criteria abc analysis based on semi-supervised clustering and explainable artificial intelligence," *IEEE Access*, vol. 11, pp. 43 778–43 792, 2023.

[81] K. L. Yung, G. T. S. Ho, Y. M. Tang, and W. H. Ip, "Inventory classification system in space mission component replenishment using multi-attribute fuzzy abc classification," *Industrial Management & Data Systems*, vol. 121, no. 3, pp. 637–656, 2021.

[82] W. Shen and Y. Qin, "Lcnet-vit-fg: a product recognition method based on the fusion of self-supervised pretrained cnn and transformer," *Journal of Electronic Imaging*, vol. 34, no. 3, pp. 033 026–033 026, 2025.

[83] K. Yang, Y. Wang, S. Fan, and A. Mosleh, "Multi-criteria spare parts classification using the deep convolutional neural network method," *Applied sciences*, vol. 11, no. 15, p. 7088, 2021.

[84] Y. Khanorkar and P. Kane, "Selective inventory classification using abc classification, multi-criteria decision making techniques, and machine learning techniques," *Materials Today: Proceedings*, vol. 72, pp. 1270–1274, 2023.

[85] J. Svoboda and S. Minner, "Tailoring inventory classification to industry applications: the benefits of understandable machine learning," *International Journal of Production Research*, vol. 60, no. 1, pp. 388–401, 2022.

[86] J. Khan, A. Ishizaka, and M. Z. Babai, "Enhancing multi-criteria inventory classification: Resolving boundary issues with vikor-fuzzy sorting," *International Journal of Production Economics*, vol. 281, p. 109526, 2025.

[87] R. Chompu-inwai and T. Thaiupathump, "Applying genetic algorithm in abc analysis for jewelry raw material inventory management," in *2023 2nd International Conference on Computer Technologies (ICCTech)*, 2023, pp. 23–28.

[88] G. C. Medina, S. E. Cabello, Y. S. Garcia, G. Quispe Santivañez, and E. Castañeda, "Inventory management model applying cyclic counting by abc-ahp classification in a laboratory equipment distribution company," in *2022 8th International Conference on Information Management (ICIM)*, 2022, pp. 151–154.

[89] I. Jackson, M. Jesus Saenz, and D. Ivanov, "From natural language to simulations: applying ai to automate simulation modelling of logistics systems," *International Journal of Production Research*, vol. 62, no. 4, pp. 1434–1457, 2024.

[90] B. Yoon, H. Kim, G. Youn, and J. Rhee, "3d position estimation of objects for inventory management automation using drones," *Applied Sciences*, vol. 13, no. 19, p. 10830, 2023.

[91] S.-C. Wu, W.-Y. Chiu, and C.-F. Wu, "Deep reinforcement learning for task assignment and shelf reallocation in smart warehouses," *IEEE Access*, vol. 12, pp. 58 915–58 926, 2024.

[92] S. Arun, S. A. Wagle, P. Nambiar, P. Panalkar, P. Ekka, V. Kumar, H. Dhiman, and A. Roy, "Optimizing micro cold storage for detecting stale food and fruits," *Science and Technology for Energy Transition*, vol. 80, p. 43, 2025.

[93] Z. Hu, S. Liu, F. Yang, X. Geng, X. Huo, and J. Liu, "Research on multi-objective optimization model of power storage materials based on nsga-ii algorithm," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, p. 76, 2024.

[94] N. Thoummala, Y. Kang, and D. Min, "A deep learning-based approach to a newsvendor problem considering uncertainty and time-varying costs," *Optimization Letters*, vol. 18, no. 9, pp. 1963–1974, 2024.

[95] I. Kaynov, M. Van Knippenberg, V. Menkovski, A. Van Breemen, and W. Van Jaarsveld, "Deep reinforcement learning for one-warehouse multi-retailer inventory management," *International journal of production economics*, vol. 267, p. 109088, 2024.

[96] S. K. Sardar, B. Sarkar, and B. Kim, "Integrating machine learning, radio frequency identification, and consignment policy for reducing unreliability in smart supply chain management," *Processes*, vol. 9, no. 2, p. 247, 2021.

[97] X. Zhu, A. Ninh, H. Zhao, and Z. Liu, "Demand forecasting with supply-chain information and machine learning: Evidence in the pharmaceutical industry," *Production and Operations Management*, vol. 30, no. 9, pp. 3231–3252, 2021.

[98] N. Mohamadi, S. T. A. Niaki, M. Taher, and A. Shavandi, "An application of deep reinforcement learning and vendor-managed inventory in perishable supply chain management," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107403, 2024.

[99] A. Adedugba and D. Inegbedion, "Inventory control optimisation: the dynamics of deterministic request model of pharmaceutical appropriation and storage," *International Journal of Operational Research*, vol. 54, no. 1, pp. 33–50, 2025.

[100] Z. Wang, W. Wang, T. Liu, J. Chang, and J. Shi, "Iot-driven dynamic replenishment of fresh produce in the presence of seasonal variations: A deep reinforcement learning approach using reward shaping," *Omega*, vol. 134, p. 103299, 2025.

[101] S. Phatangare, L. Sharma, A. Mate, O. Jaiswal, and D. Pardeshi, "Revolutionizing food ordering: An ml-driven system for predictive sales and inventory optimization," in *2025 International Conference on Computing Technologies (ICOCT)*. IEEE, 2025, pp. 1–6.

APPENDIX

As shown in Table V, the acronyms and full names of various AI methods are presented.

TABLE V. TABLE OF AI ALGORITHM ACRONYMS

Acronym	Complete name
AHK-ELM	Attention-based Hybrid Kernel Extreme Learning Machine
ANN	Artificial Neural Network
BiLSTM	Bi-directional Long Short-Term Memory
BNN	Bayesian Neural Network
DL	Deep Learning
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
DRLOM	Deep Reinforcement Learning-based Ordering Mechanism
DT	Decision Tree
DTL	Decision Tree Classifier
ES	Exponential Smoothing
GA	Genetic Algorithm
GB	Gradient Boosting
GBDT	Gradient Boosting Decision Trees
IMOHHO	Improved Multi-Objective Harris Hawk Optimization
LGB	Light Gradient Boosting Machine
LSTM	Long Short-Term Memory
MAML	Model-Agnostic Meta-Learning
MARL	Multi-Agent Reinforcement Learning
MCTS	Monte Carlo Tree Search
ML	Machine Learning
MLR	Multiple Linear Regression
MLR	Natural Language Processing
MNB	Multinomial Naive Bayes
MNLR	Multinomial Logistic Regression
NARXNN	Nonlinear Autoregressive Exogenous Neural Network
PSO	Particle Swarm Optimization
RF	Random Forest
SAC	Soft Actor-Critic
SHAP	Shapley Additive Explanations
SVM/R	Support Vector Machine/Regression

Glioma Classification Using Harris Hawks-Driven Optimized Gradient Boosting Classifier Along with SHAP-Based Interpretability

SM Naim[1], Jun-Jiat Tiang[2]*, Abdullah-Al Nahid[3]*
Electronics and Communication Engineering Discipline, Khulna University, Khulna 9208, Bangladesh[1,3]
Centre for Wireless Technology-Centre of Excellence for Intelligent Network-Faculty of Artificial Intelligence and Engineering, Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia[2]

***Abstract*—Gliomas are considered one of the most lethal and aggressive types of brain cancer, responsible for countless deaths worldwide. This study seeks to improve glioma classification using cutting-edge machine learning (ML) techniques to differentiate between glioma subtypes based on clinical and genomic data. The goal is to identify important biomarkers and features influencing glioma classification, with an emphasis on improving feature selection and model interpretability. For glioma classification, the Gradient Boosting Classifier (GBC) was employed. The Harris Hawks Optimization (HHO) algorithm was used for feature selection and hyperparameter fine-tuning to enhance the model's performance. Additionally, SHapley Additive exPlanations (SHAP) were applied to improve model interpretability and to better understand feature contributions.The Gradient Boosting (GB) method yielded the best performance among the selected models, achieving an accuracy of 88.40%, precision of 87.3%, recall of 88.48%, and an F1 score of 88.29%, with feature selection and hyperparameter tuning using the Harris Hawks Optimization. These results highlight the significance of hyperparameter tuning and feature selection in enhancing classification performance. Key features such as IDH1, Age at Diagnosis, and EGFR were identified as the most influential in distinguishing glioma subtypes. SHAP analysis further confirmed the importance of these features in the model.This study shows that the Gradient Boosting Classifier (GBC), optimized with Harris Hawks Optimization (HHO), significantly improves glioma classification, achieving a high F1 score. Key features like IDH1, Age at Diagnosis, and EGFR were identified, showcasing its potential for enhanced glioma diagnosis.**

***Keywords*—Glioma; gradient boosting; Harris Hawks Optimization (HHO); SHAP; feature selection; interpretability; TCGA; IDH1; EGFR**

I. Introduction

Brain and nervous system cancer is one of the most dangerous cancers in today's age. The death toll is rising day by day. According to research by the World Health Organization, there were approximately 251,000 deaths from brain and nervous system cancer in 2020, across both sexes, up to 85 years of age [1]. Among these, gliomas represent a diverse group of tumours originating from glial cells in the brain and pose significant diagnostic and therapeutic challenges in neuro-oncology [2]. Accurate grading of gliomas is critical, as it informs clinical decision-making and influences patient outcomes. However, conventional grading methods often lack precision and are resource-intensive. As illustrated in Fig. 1, the global burden motivates the need for reliable, explainable decision support.

Survival rates for glioma patients remain low, especially for high-grade variants such as glioblastoma multiforme (GBM), which alone accounts for more than 60 percent of adult primary brain tumours [6], [7]. The 2007 WHO classification introduced a system based on histological features, which was later enhanced by incorporating molecular markers such as IDH mutations [4]. In recent years, the convergence of neuroinformatics, computational neuroimaging, and machine learning has provided promising avenues for improving glioma classification. Machine learning algorithms can analyse complex, high-dimensional data from genomic and clinical sources to uncover patterns that traditional methods may overlook. One of the key challenges in glioma grading is the reliance on expensive molecular tests like IDH1/IDH2 mutation analysis, which can cost between USD 135–1800 and take up to two weeks to process [9], [10]. Additionally, variables such as age, sex, and clinical symptoms also influence tumour behaviour but are not always integrated effectively due to limited dataset annotations [11].

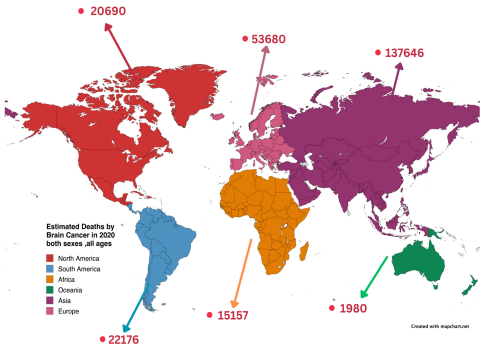


Fig. 1. Estimated number of brain cancer deaths by region in 2020.

Gliomas are among the most prevalent and deadly forms of brain cancer [3]. Early and accurate detection is crucial for prognosis and treatment decisions. Machine learning has shown great promise in extracting valuable insights from complex medical datasets, making it possible to predict disease

*Corresponding author.

outcomes with higher accuracy. This paper leverages advanced techniques such as Gradient Boosting Classifiers (GBC) and optimisation algorithms like HHO to identify key features that improve glioma diagnosis, with a specific focus on their interpretation through SHAP analysis [21].

Despite many promising studies, important gaps remain. A number of works maximize accuracy without explaining the predictions; others provide explanations but rely on only one data type (for example, imaging alone) or evaluate models using a single cross-validation loop that can overestimate performance. It is still uncommon to jointly perform feature selection and hyperparameter tuning on combined clinical+genomic predictors, then audit the final model with SHAP and report fold-aggregate results that better reflect generalization and clinical use.

This paper addresses these gaps with a simple, end-to-end pipeline that uses HHO with a GBC to jointly pick features and tune hyperparameters on TCGA-LGG/GBM data, followed by SHAP to check that the most influential variables—such as *IDH1*, *EGFR*, *TP53*, *NF1*, and Age—agree with contemporary glioma biology and the WHO-2021 system [5] [22]. We evaluate the model with cross-validation and summarize results across folds rather than relying on a single split. We also explain why gains in threshold-based metrics (accuracy, F1) can coexist with small changes in ROC AUC, and we outline how the outputs could support care by triaging cases for confirmatory molecular testing and by providing feature-level explanations suitable for tumor-board discussion.

II. Related Works

TCGA (The Cancer Genome Atlas) data have been used extensively to classify gliomas [8]. To differentiate glioblastoma multiforme (GBM) patients from others, Ko and Brody, for instance, used a gradient boosting classifier on TCGA copy-number data, obtaining an AUC of 0.875 [13]. Likewise, Sánchez-Marqués et al. demonstrated that ensemble approaches (e.g., CatBoost) outperformed standard classifiers on TCGA glioma cohorts by using TCGA-LGG and TCGA-GBM molecular and clinical data for glioma grade prediction [14]. These investigations emphasize how rich the TCGA-LGG/GBM dataset—which consists of hundreds of patients and clinical/molecular data—is for glioma subtype and grade classification problems. Underlining the relevance of TCGA in glioma research, other studies have built predictive models of glioma subtype or prognosis using TCGA-derived data (gene expression, mutations).

Medical and glioma classification has seen especially successful application for gradient boosting techniques. In oncology classification tasks, ensemble tree-boosting algorithms including XGBoost, LightGBM, and CatBoost routinely produce state-of-the-art results. Tang et al. used XGBoost on TCGA GBM transcriptomic profiles, for example, to classify GBM into its three subtypes (proneural, classical, mesenchymal), identifying a five-gene signature and approximating 80% accuracy [17]. Xia et al. trained GradientBoost and LightGBM models on MRI radiomic features to differentiate GBM from solitary metastases, achieving ROC-AUC > 0.90, and used SHAP to interpret feature contributions [16]. Gradient boosting approaches have also performed powerfully in radiomics-based brain tumor studies. Using MRI data, Kha et al. built an XGBoost model to forecast 1p/19q co-deletion in lower-grade glioma and used SHAP to choose the most useful radiomic features [15]. These and other studies—e.g., Sánchez-Marqués et al. using CatBoost [14]—show that gradient boosting classifiers are well-suited to complex, high-dimensional medical data, producing high accuracy in glioma grading and biomarker prediction.

Applied for feature selection and hyperparameter tuning in biomedical machine learning, Harris Hawks Optimization (HHO) is a new nature-inspired metaheuristic. Originally proposed in 2019 by Heidari et al., HHO has been modified for high-dimensional biomedical data in later work. Elgamal et al., for instance, chose ideal feature subsets in medical datasets by means of an enhanced HHO coupled with simulated annealing. Pirgazi et al. proposed a two-stage filter-wrapper approach whereby an enhanced HHO-based wrapper (with GRASP) further refines features chosen by a filter to identify an optimal subset for classification [21]. Apart from feature selection, HHO has been applied effectively to adjust model hyperparameters in clinical predictive models. Using HHO in a COVID-19 detection system, Kumar et al. optimized the hyperparameters of boosting classifiers (XGBoost, LightGBM, etc.), improving model performance and enabling integration of SHAP for feature analysis [18]. In a liver cirrhosis prediction task, Nalasari et al. similarly combined XGBoost with HHO for hyperparameter tuning, reporting notably better accuracy and less overfitting than standard XGBoost [19]. In the neuro-oncology setting [12], Kurdi et al. embedded HHO into a convolutional neural network (HHOCNN) for MRI-based brain tumor detection; their HHO-optimized CNN attained 98% accuracy on benchmark MRI datasets [20]. These cases show that by automatically selecting features or tuning parameters, HHO can efficiently improve model training; moreover, its application to biomedical classification is becoming rather common.

Rising in prominence in healthcare, interpretable artificial intelligence techniques such as SHapley Additive exPlanations (SHAP) offer insights into black-box model decisions. Several glioma research studies have used SHAP to clarify predictive characteristics. Before training the XGBoost model for 1p/19q status, Kha et al. ranked radiomic features using SHAP values, so increasing the transparency of the model [15]. Xia et al. calculated mean SHAP values for radiomic features in imaging studies to underline the most important predictors of tumor type [16]. Using SHAP to offer both global and local explanations of model outputs in a glioma survival-prediction web tool helped to show how clinical and molecular factors influence risk projections. Likewise, Kumar et al. used SHAP analysis on their COVID-19 boosting model to find important risk factors and simplify the predictions [18]. These papers demonstrate that SHAP is a useful tool in medical machine learning since it quantifies feature importance in complicated models, thus improving clinician confidence. SHAP has been applied not only for MRI/radiomics models but also for genomic and clinical models in the glioma domain, exposing which genetic changes or patient factors drive predictions.

All told, previous studies have shown the value of TCGA-based ML for glioma, the efficacy of gradient-boosting classifiers in such tasks, the promise of HHO for optimizing feature

sets and model parameters, and the importance of SHAP in model interpretation. These components have, nevertheless, mainly been studied separately. The present work validates the model with SHAP-based interpretability and combines these developments by using HHO to execute both feature selection and hyperparameter tuning of a gradient boosting classifier on TCGA glioma data. We know of no prior reports of this combined approach: an HHO-optimized GBC trained on TCGA genomic data using SHAP validation. It jointly improves model performance and transparency, bridging gaps in current work. While prior studies have shown that gradient boosting can achieve strong accuracy, many do not explain predictions in a way that is practical for clinical discussion, and others rely on a single data type rather than integrating clinical and genomic information. Several works also tune models within a single cross-validation loop, which can make the reported numbers look stronger than they are in routine use. As a result, three limitations remain common: modest gains in accuracy without clear interpretability, limited use of combined clinical+genomic inputs, and evaluation protocols that do not emphasize generalization. Our approach is designed to address these points by jointly selecting features and tuning hyperparameters with HHO, auditing the final model with SHAP, and reporting fold-level summaries that reflect how the model is likely to behave beyond one split.

III. Methodology

Based on the dataset, this work has carried out a classification task to distinguish between patients with gliomas and those without. Since the dataset is labeled, we have thought of using the supervised procedure to address this issue. Along with task classification, we have also carried out some DAT research, feature selection, and feature prioritization. We have used the SHAP method, a subset of the explainable AI (XAI) approach, along with a few filter techniques for feature selection. One could think of the XAI as a wrapper-based framework. The end-to-end pipeline—preprocessing, SMOTE, HHO-based feature selection and tuning, model training, and evaluation—is summarized in Fig. 2.

A. Dataset and Preprocessing

We conducted all experiments using a dataset obtained from the well-known and publicly accessible repository of genome atlas data on TCGA . The dataset was created using information from the TCGA-LGG and TCGA-GBM projects. It includes three clinical factors—Gender, Age at diagnosis, and Race—along with 20 commonly mutated molecular biomarkers, all gathered from 839 patients diagnosed with LGG or GBM. Looking at Table I, we can see that all the predictors are categorical, with the exception of the Age at diagnosis, which is represented as a numerical value.

The molecular characteristics are indicated by values of 0 for not mutated and 1 for mutated, based on the TCGA case number. It's important to highlight that there was no need to use any deletion or imputation techniques, as the dataset utilized in the experiments was complete, with no missing values in any of the attributes (predictor variables).

The dataset contains 24 attributes of 839 patients. Among them, 352 are Glioma patients and 487 are non-Glioma patients. As the data was imbalanced, we balanced the data using the SMOTE function.

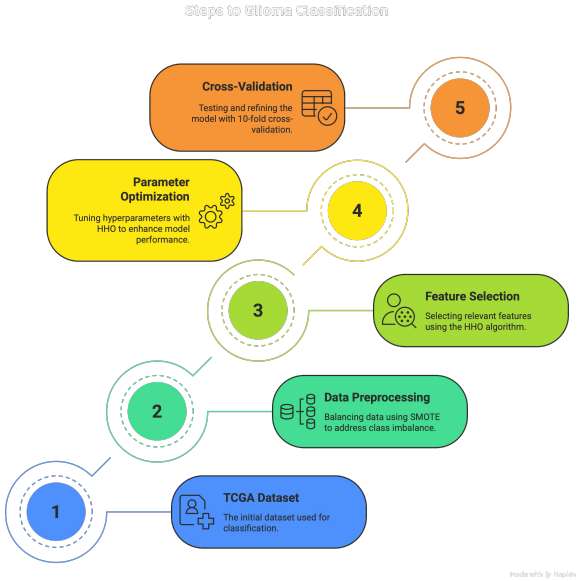


Fig. 2. Methodology for glioma classification using the TCGA dataset. This diagram illustrates the five key steps: dataset selection, data preprocessing with SMOTE, feature selection using HHO, hyperparameter optimization with HHO, and model testing via 10-fold cross-validation.

TABLE I. Predictors in the Dataset Used for Classification

Predictor Name	Category	Possible Values
Gender	Clinical	0, 1
Age	Clinical	[14.42, 89.29]
Race	Clinical	0, 1, 2, 3
IDH1	Molecular	0, 1
TP53	Molecular	0, 1
ATRX	Molecular	0, 1
PTEN	Molecular	0, 1
EGFR	Molecular	0, 1
CIC	Molecular	0, 1
MUC16	Molecular	0, 1
PIK3CA	Molecular	0, 1
NF1	Molecular	0, 1
PIK3R1	Molecular	0, 1
FUBP1	Molecular	0, 1
RB1	Molecular	0, 1
NOTCH1	Molecular	0, 1
BCOR	Molecular	0, 1
CSMD3	Molecular	0, 1
SMARCA4	Molecular	0, 1
GRIN2A	Molecular	0, 1
IDH2	Molecular	0, 1
FAT4	Molecular	0, 1
PDGFRA	Molecular	0, 1

To better understand the interdependence among the features, we calculated the Pearson correlation matrix for all numerical variables in the dataset. The correlation matrix above (Fig. 3) represents, using color and circle size, the effect and direction of linear relationships between features. The blue circles are for negative correlations, the red circles are for positive correlations, and the larger the circle, the stronger the relationship.

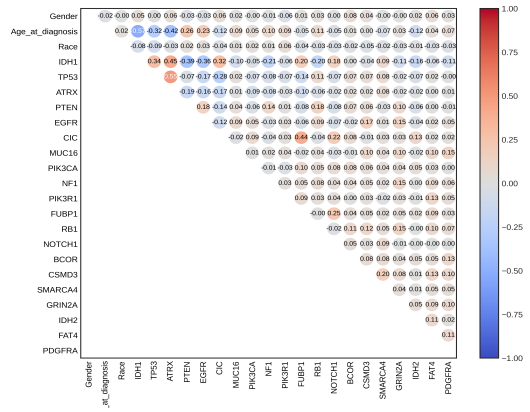


Fig. 3. Upper-triangle correlation matrix of selected clinical and genomic features. Circle size and color intensity indicate the strength and direction of Pearson correlation coefficients (red = positive, blue = negative).

An upper-triangle form such as this one draw attention to potential multicollinearity and redundancies. Specifically, TP53 and IDH1, and Age at diagnosis attributes are moderately correlated ($r2 \geq 0.1$) with multiple other attributes, and therefore may also be interesting predictive features. On the other hand, the lack of strong correlation between most features seems to suggest that this dataset can be used for evaluating and optimizing feature selection separately.

B. Visualization of Data Distribution Using t-SNE

In order to gain insights into the structure of our gene expression dataset and the distribution of our classes within it, we performed dimensionality reduction using t-distributed Stochastic Neighbor Embedding (t-SNE). And this is a nonlinear method which embeds the high-dimensional features into a 2D space preserving local similarities so it is very useful for visualizing patterns and class separability. Fig. 4: t-SNE projection of the entire dataset showing class-wise distribution.

Fig. 4. T-SNE projection of the entire dataset showing class-wise distribution.

In Fig. 4, we can see the t-SNE projection of the entire dataset showing Class 0 (purple) and Class 1 (orange) samples in separate, though partially overlapping clusters. This indicates that the input space in high-dimensional has an underlying structure that is probably separable by the proper classifier. This consistency in data distribution across the full dataset validates the efficiency of our preprocessing pipeline (as the distribution of pre-processed data in each set is analogous), and these visualizations make it all the more reassuring.

C. Gradient Boosting Classifier

In this subsection, we discuss the Gradient Boosting classifier used to classify glioblastoma. A classifier is a type of machine learning (ML) algorithm that categorizes data into predefined classes. In this case, the classifier uses patient characteristics, clinical data, to determine whether or not the patient has glioblastoma.

Gradient Boosting (GB) is an advanced ensemble learning algorithm that sequentially constructs weak learners (typically decision trees) to form a strong classifier. It minimizes the prediction error by iteratively adding models that correct the residuals of the previous ensemble. The objective function for GB is:

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{1}$$

Where:

- $L(\theta)$ is the objective function representing the overall loss.
- $l(y_i, \hat{y}_i)$ is the loss function measuring prediction error for each data point.
- $\Omega(f_k)$ is the regularization term to penalize model complexity.

The regularization function is defined as:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2}$$

Where:

- T is the number of leaves in the tree.
- w_j is the weight of the j^{th} leaf.
- γ, λ are regularization hyperparameters that control the strength of the regularization.

The model is updated at each iteration using:

$$F_t(x) = F_{t-1}(x) + \alpha h_t(x) \tag{3}$$

Where:

- $F_t(x)$ is the model's prediction after t iterations.
- α is the learning rate controlling the size of updates.
- $h_t(x)$ is the weak learner at iteration t.

D. Performance Metrics

To evaluate the classification performance, several standard metrics were used. These metrics provide insights into the model's ability to correctly identify and classify instances. The key metrics considered include:

- Accuracy: This metric measures the proportion of correctly classified instances out of the total instances. It provides an overall effectiveness of the classification model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

- Precision: Precision quantifies the accuracy of the positive predictions made by the model. It is the ratio of true positives to all instances classified as positive, showing how many of the positive predictions were correct.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$

- Recall: Recall, also known as sensitivity, measures the model's ability to correctly identify all relevant positive instances. It is the ratio of true positives to the total number of actual positives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6}$$

- F1 Score: The F1 Score is the harmonic mean of Precision and Recall, providing a single metric that balances both concerns. A higher F1 score indicates a better balance between Precision and Recall.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Where:

- TP = True Positive
- TN = True Negative
- FP = False Positive
- FN = False Negative

Additionally, Confusion Matrices and Receiver Operating Characteristic (ROC) Curves were used for visual interpretation of classification performance. The confusion matrix allows for a detailed breakdown of the true vs. predicted classifications, while ROC curves provide insights into the trade-off between the True Positive Rate (Recall) and False Positive Rate at various thresholds.

These metrics offer a comprehensive understanding of the model's classification capabilities and its strengths and weaknesses in distinguishing between classes.

E. Feature Selection Using Harris Hawk Optimization (HHO) Algorithm

Feature selection is the process of finding a noise-free, effective set of features from a given dataset that improves the model's performance on that dataset. The feature selection approach can be broadly divided into three main categories: embedded method, filter-based method, and wrapper-based method. In the wrapper-based method, a subset of the total features is evaluated using a machine learning (ML) algorithm iteratively to find the best feature subset. Metaheuristic algorithms are mostly used in this case to iteratively find an optimum feature subset that provides maximum performance.

To optimize feature selection, we employed the Harris Hawk Optimization (HHO) algorithm—a metaheuristic inspired by the cooperative hunting strategy of Harris hawks. The position of each hawk represents a candidate feature subset. The optimization is driven by exploration and exploitation phases:

- Exploration: Hawks randomly search for solutions based on prey energy.
- Exploitation: Attack strategies like soft besiege, hard besiege, and sudden dives are applied based on the prey's escaping energy.

The energy of the prey is given by:

$$E = 2E_0 \left(1 - \frac{t}{T}\right) \tag{8}$$

Where:

- E_0 is the initial energy of the prey,
- t is the current iteration,
- T is the maximum number of iterations.

Hawks update their position based on the escape energy E using one of the following position update strategies:

$$X(t+1) = X_{\text{prey}}(t) - E \cdot |J \cdot X_{\text{prey}}(t) - X(t)| \tag{9}$$

Where:

- $X(t)$ is the current hawk position,
- $X_{\text{prey}}(t)$ is the position of the prey (best solution),
- J is the random jump strength factor.

For the optimization process, we used the F1 score as the cost function to evaluate the performance of each feature subset. The F1 score, which is the harmonic mean of precision and

recall, was chosen because it provides a balanced evaluation of model performance, especially in imbalanced datasets. The F1 score was computed for each subset, and the feature subset that yielded the highest F1 score was selected as the optimal subset.

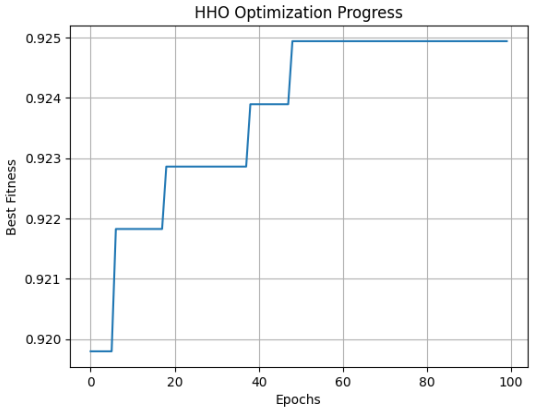


Fig. 5. Harris Hawks Optimization (HHO) convergence curve showing the improvement of best fitness over 100 epochs. The figure illustrates a typical step-wise convergence pattern characteristic of HHO's exploitation-exploration balance.

The Harris Hawks Optimization (HHO) algorithm progressively improves the model's F1 score by selecting optimal feature subsets. As shown in Fig. 5, the F1 score increases in a step-wise manner over the epochs. This reflects HHO's balance between exploration and exploitation, leading to a near-optimal solution.

We initialized 30 hawks over 50 iterations and evaluated subsets using the Gradient Boosting Classifier with 100 estimators. An elitism strategy was used to preserve the top 2 performing hawks during each iteration. The F1 score was maximized by iteratively adjusting the feature set and hyperparameters, ensuring that the final selected features and parameters provided the highest classification performance.

We applied the Harris Hawks Optimization (HHO) algorithm (Algorithm 1) to select an optimal subset of features, which reduced the feature space and improved classification accuracy. The solution sets provided by HHO indicate which features to select. These subsets are then used in the Gradient Boosting Classifier (GBC) for classification, with the F1 score serving as the fitness value for each agent (feature subset). The F1 score effectively balances precision and recall, making it a robust performance metric. In this process, we utilized 30 hawks over 50 iterations. The HHO algorithm updates the hawks' positions based on the best fitness value found at each step. By the end of the process, HHO selects the optimal feature subset, ensuring that the classification model is both efficient and accurate.

Algorithm 1 Feature Selection Using HHO for GBC

Initialize the hawk population $X(i)$ for $i = 1, 2, \ldots, n$
for each hawk in the population **do**
    Decode the feature set from the hawk position
    Calculate the fitness value (F1 score) using the decoded feature set with the Gradient Boosting classifier
end for
Compare the fitness values of each hawk and set X^* as the best hawk
while $t <$ maximum iteration number **do**
    for each hawk in the population **do**
        Update parameters
        if $r < 0.5$ **then**
            if $|A| < 1$ **then**
                Update hawk position towards the global best X^* using the exploration phase
            else
                Select a random position X_{rand}
                Update hawk position towards the random position using the exploration phase
            end if
        else
            Update hawk position towards the global best X^* using the exploitation phase
        end if
    end for
    for each hawk in the population **do**
        Decode the feature set from the hawk position
        Calculate the fitness value (F1 score) using the decoded feature set with the Gradient Boosting classifier
    end for
    Compare the fitness values of each hawk
    if a better solution (feature set) is found **then**
        Update X^*
    end if
end while
Save the best solution set as the final feature set

IV. Results

A. Feature Selection and Hyperparameter Configuration

The feature selection process resulted in the identification of a subset of features that contributed most significantly to the classification performance. The selected features are shown in Table II.

TABLE II. Selected Features After Harris Hawks Optimization

Selected Feature Name			
Gender	Age_at_diagnosis	Race	IDH1
ATRX	PTEN	EGFR	PIK3CA
NF1	SMARCA4	IDH2	PDGFRA

These features were selected based on their ability to differentiate between glioma subtypes. The selection process improved the model's efficiency by reducing the dimensionality of the input data.

The optimal hyperparameters for the Gradient Boosting Classifier (GBC) were determined through the Harris Hawks Optimization (HHO) process. The configurations are shown in Table III.

TABLE III. OPTIMIZED HYPERPARAMETERS FOR THE GBC MODEL

Hyperparameter	Original Configuration	Hyperparameter Tuning	Feature Selection & Tuning
n_estimators	100	144	150
learning_rate	0.1	0.0632	0.0139
max_depth	3	5	5
subsample	1.0	0.8153	0.8465

The optimized values for the hyperparameters show a clear reduction in the learning rate and an increase in the number of estimators, which contribute to the improved performance of the model.

B. Model Performance Comparison

The performance of the model was assessed across three different scenarios: using the original dataset, after hyperparameter tuning, and after both feature selection and hyperparameter tuning. The results, measured using various performance metrics, are summarized in Table IV.

TABLE IV. PERFORMANCE COMPARISON ACROSS DIFFERENT CONFIGURATIONS

Metric	Original Dataset	Hyperparameter Tuning	Feature Selection & Tuning
Accuracy	0.8809 ± 0.0269	0.8830 ± 0.0305	0.8840 ± 0.0266
Precision (Macro)	0.8834 ± 0.0272	0.8846 ± 0.0316	0.8873 ± 0.0264
Recall (Macro)	0.8812 ± 0.0257	0.8826 ± 0.0297	0.8848 ± 0.0253
F1 Score (Macro)	0.8796 ± 0.0265	0.8816 ± 0.0303	0.8829 ± 0.0260
ROC AUC	0.9240 ± 0.0288	0.9304 ± 0.0255	0.9269 ± 0.0315

The results show that feature selection combined with hyperparameter tuning achieved the highest scores across all evaluation metrics. Specifically, the accuracy improved from 0.8809 to 0.8840, while the ROC AUC decreased slightly from 0.9304 to 0.9269 compared to hyperparameter tuning alone, highlighting the effectiveness of feature selection in enhancing model performance. The improvements demonstrate that feature selection, when coupled with hyperparameter tuning, can optimize model performance, confirming its applicability in resource-limited environments.

C. ROC Curve for the Three Models

The Receiver Operating Characteristic (ROC) curve for the best-performing models (Feature Selection & Hyperparameter Tuning, Original Dataset, and HHO Optimization) is shown in Fig. 6. The curves demonstrate the performance of each model, with their respective Area Under the Curve (AUC) values.

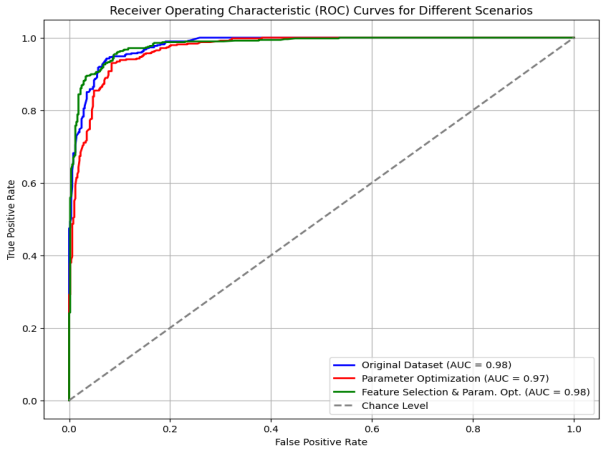


Fig. 6. ROC curve comparison for all three models.

The feature selection + tuning configuration delivers higher accuracy and F1 at the working threshold, while its ROC AUC is slightly lower than the tuning-only configuration. This pattern is consistent with optimizing for F1 rather than global ranking and matches the values reported in Table IV.

D. Confusion Matrix Across Three Models

The confusion matrices for the best fold from three different models—Original Feature Set & Hyperparameter Tuning (OFS & HT), Original Dataset, and Selected Feature Set (SFS) & Hyperparameter Tuning—are presented in Fig. 7a, 7b, and 7c, respectively. These matrices display the key classification metrics: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each model. Below are the interpretations drawn from these confusion matrices:

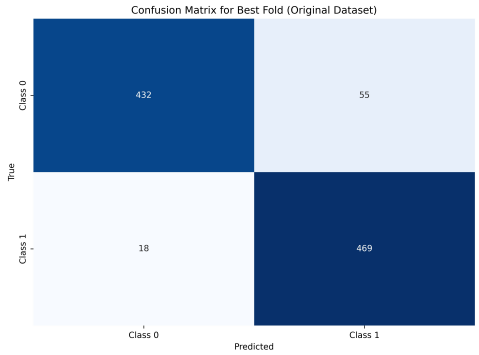
The confusion matrix for the Original Dataset model performs reasonably well, but it reveals a higher number of false positives and false negatives compared to the OFS & HT model. This suggests that, while this model achieves good accuracy, it would benefit from the enhancements offered by feature selection and hyperparameter tuning.

In summary, all models performed well, but the Feature Selection & Hyperparameter Tuning (OFS & HT) model demonstrated the best performance with the highest number of correct classifications. This highlights the importance of feature selection and hyperparameter tuning in achieving optimal classification results.

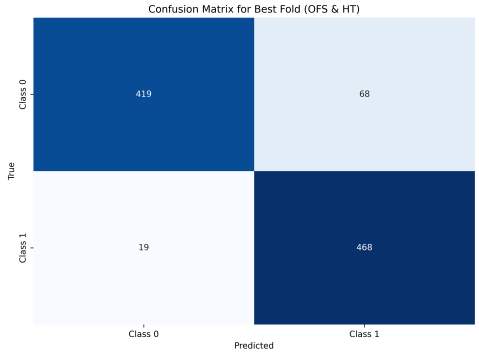
E. K-Fold Cross-Validation Results Across Three Models

Fig. 8 presents the performance metrics across 10-fold cross-validation for each of the three models. Boxplots show the spread and central tendency of each metric, including accuracy, precision, recall, F1 score, and ROC AUC, providing a comparison of model stability and performance.

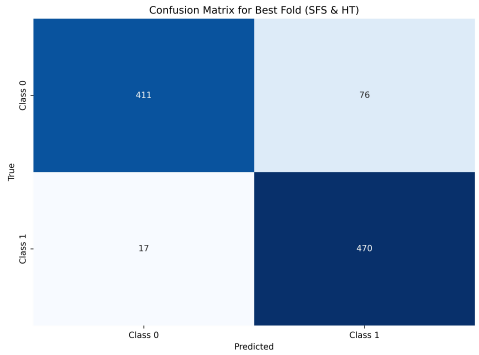
These boxplots highlight that the SFS & HT model is the best-performing model, exhibiting consistent and superior performance across all evaluation metrics.



(a) Confusion matrix for original dataset.



(b) Confusion matrix for hyperparameter tuning only.

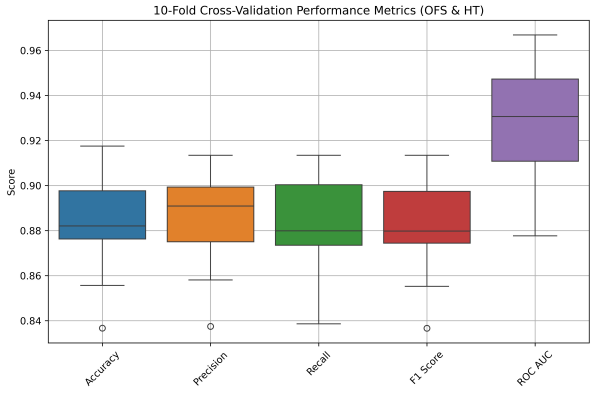


(c) Confusion matrix for feature selection + tuning.

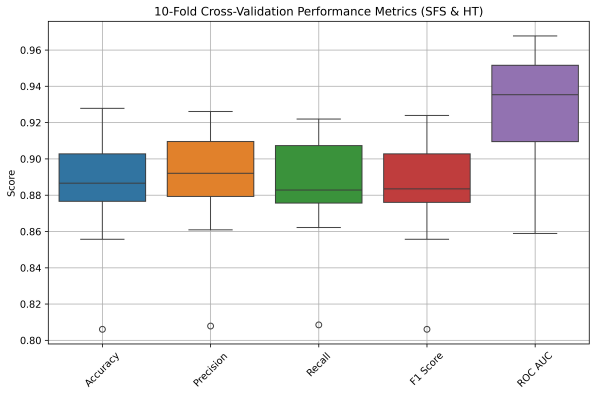
Fig. 7. Confusion matrices for the best fold across three models: Original dataset, hyperparameter tuning only, and feature selection + tuning.

F. SHAP-Based Validation of Selected Features

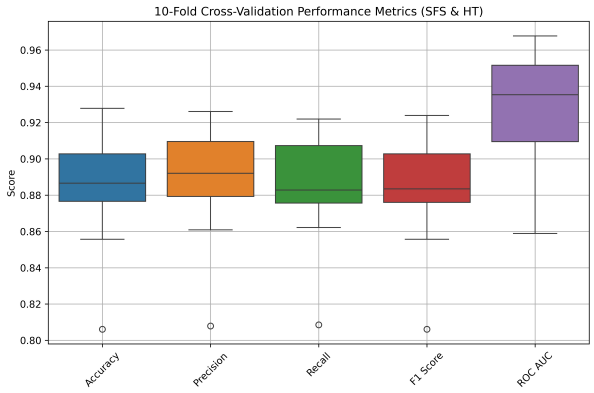
We used SHAP (SHapley Additive exPlanations) to evaluate the output of the trained model on the training data, thereby validating the resilience of the feature subset chosen using the Harris Hawks Optimization (HHO) algorithm. SHAP, based on cooperative game theory, offers consistent, locally accurate feature attributions, making it a useful instrument for post-hoc model interpretability. By providing a clear understanding of how each feature contributes to model predictions, SHAP helps validate the feature selection process and ensures the reliability of the model's decisions.



(a) Original dataset.



(b) Hyperparameter tuning.



(c) Feature selection + tuning.

Fig. 8. 10-Fold cross-validation performance metrics for the three experimental models: Original dataset, hyperparameter tuning only, and feature selection + hyperparameter tuning. Metrics include accuracy, precision, recall, F1 score, and ROC AUC.

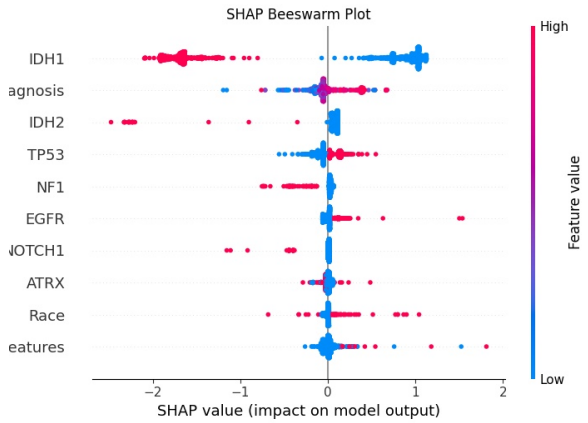


Fig. 9. SHAP summary plot illustrating the impact of individual features on the model's output for the training set. Red and blue represent high and low feature values, respectively.

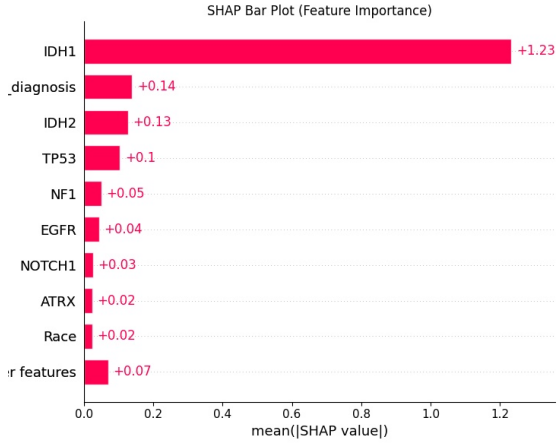


Fig. 10. Mean absolute SHAP values indicating the overall importance of features across all training samples. Top-ranked features include IDH1, Age_at_diagnosis, PTEN, and IDH2.

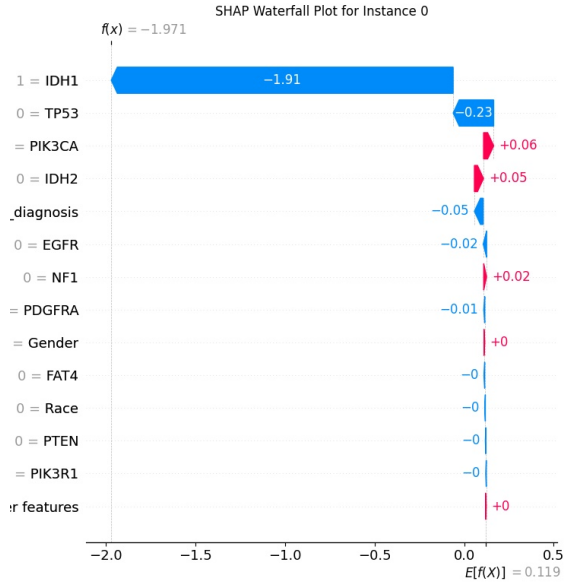


Fig. 11. SHAP waterfall plot illustrating individual feature contributions for a specific prediction. The plot shows the cumulative impact of each feature.

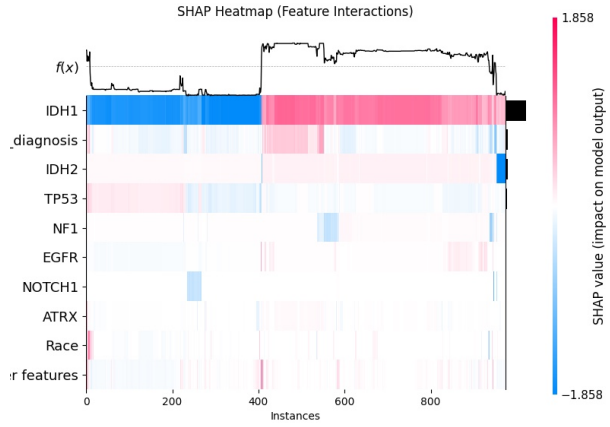


Fig. 12. SHAP heatmap showing feature importance and interactions across all predictions.

Fig. 9 illustrates the SHAP summary plot, which shows the impact of individual features on the model's output for the training set. Red and blue represent high and low feature values, respectively. The most powerful features influencing model predictions are revealed by the SHAP summary plot and the bar plot of mean absolute SHAP values (Fig. 10). Not least among these are IDH1, Age_at_diagnosis, PTEN, IDH2, EGFR, TP53, and NF1. These traits have great predictive value since they show as high-impact contributors over several samples.

The SHAP waterfall plot in Fig. 11 illustrates individual feature contributions for a specific prediction. The plot displays the cumulative impact of each feature, helping us understand how the model arrived at its output for a particular instance.

Finally, Fig. 12 shows the SHAP heatmap, which illustrates feature importance and interactions across all predictions. It provides a visual representation of how features interact and their relative importance in the overall model.

Especially, several of the top SHAP-ranked features are included in the subset of features chosen by the HHO algorithm — `['Age_at_diagnosis', 'IDH1', 'EGFR', 'NF1', 'PIK3R1', 'FUBP1', 'NOTCH1', 'BCOR', 'IDH2', 'FAT4']` —. Particularly for IDH1, Age_at_diagnosis, EGFR, NF1, IDH2, NOTCH1, and BCOR, the overlap between the SHAP analysis and HHO selection offers independent proof that the optimization procedure effectively found highly impactful factors for the model.

This agreement suggests that the HHO-based feature selection technique captured features of actual relevance for classification, thereby supporting the validity of this approach. The general interpretability and resilience of the proposed framework are supported by the agreement between model explanation and optimization results.

We tuned the model to work best at one practical decision point (the F1 balance), not across every possible threshold. That is why accuracy and F1 increased while ROC AUC, which reflects overall ranking across all thresholds, dipped a little: some score ordering changed, but decisions at the chosen cutoff improved. The SHAP results show that a small set of biologically sensible features—IDH1, EGFR, TP53, NF1, and Age—push predictions in expected directions; "high-risk" patterns raise the score, while "lower-risk" patterns reduce it. In practice, a high score with a few strong drivers can prompt faster confirmatory testing or earlier team review, whereas a low score with mixed or weak drivers can follow routine care. This ties the numbers back to clinical meaning and explains why the confusion matrices show fewer mistakes at the working threshold.

V. Discussion

This study shows that combining Harris Hawks Optimization (HHO) with a Gradient Boosting Classifier (GBC) can produce a compact and accurate model for glioma classification while keeping the reasoning behind predictions understandable. In our view, the most encouraging aspect is that the features highlighted by the model—such as IDH1, EGFR, TP53, NF1, and Age—are consistent with the biological story clinicians expect to see in glioma. This agreement suggests the model is learning meaningful patterns rather than overfitting to noise, and it makes the outputs easier to discuss in multidisciplinary settings.

We also observe a clear trade-off between threshold-based metrics and ranking-based metrics. After feature selection and tuning, accuracy and F1 improved, while ROC AUC decreased slightly. This is a reasonable outcome because the optimization process concentrated on improving the balance of precision and recall at a working threshold, not on maximizing performance across all possible thresholds. In practical terms, the model became better at identifying true cases at an operating point that matters for care, even if overall ranking changed a little. Because clinical use often depends on a specific decision threshold, we consider this an acceptable and transparent trade-off.

Another practical strength is parsimony. The HHO search consistently pushed the model toward a smaller set of informative variables, which reduces redundancy, speeds up training and inference, and lowers the cost of deployment. The resulting explanations from SHAP are easier to interpret at the patient level because fewer features dominate the prediction. In a clinical workflow, this can help frame conversations such as why the model flagged a case and which factors were most influential at that moment.

Finally, we see a realistic path to use: align the decision threshold with local practice, check probability calibration, review explanations alongside routine clinical information, and monitor performance over time. The goal is not to replace clinical judgment but to support it with a tool that is fast, consistent, and explainable.

VI. Limitations

This work uses a single public dataset for development and internal testing. Although it includes many patients, it may not capture the full diversity of real-world populations, imaging or sequencing methods, and clinical practice. As a result, the model's performance could change when applied to other hospitals or regions. External validation on independent cohorts from different institutions is needed to confirm generalizability.

The labeling scheme and endpoint definition may not fully match current clinical reporting standards. While the model distinguishes classes effectively within the dataset, future versions should align labels more closely with contemporary diagnostic categories to better reflect how decisions are made in practice.

Class imbalance was addressed with synthetic oversampling during training, which can influence the distribution and calibration of predicted probabilities. Although this helped improve recall and F1 at the chosen threshold, probability estimates may require calibration before the model is used to trigger clinical actions. Threshold choice should also reflect local risk tolerance and downstream resource constraints.

Metaheuristic optimization introduces randomness through initialization and fold splits. Although we used cross-validation and stable settings, minor variation between runs is expected. A more exhaustive stability assessment across multiple seeds and repeated folds would provide stronger evidence that the selected feature set and hyperparameters are robust.

Lastly, while SHAP improves transparency, it is a post-hoc explanation technique and does not prove causality. Some interactions between features may be complex and context-dependent, and explanations should be interpreted as supportive evidence rather than definitive biological mechanisms. Future work will focus on external testing, probability calibration, threshold setting with clinical input, and prospective evaluation in a workflow that tracks utility and safety over time.

VII. Conclusions

This study demonstrates that the Gradient Boosting Classifier (GBC), optimized using Harris Hawks Optimization (HHO) for feature selection and hyperparameter tuning, significantly enhances glioma classification performance. Despite reducing the feature set by 10 features, the Feature Selection & Hyperparameter Tuning (OFS & HT) model still outperforms other models in terms of accuracy, precision, recall, and F1 score. By selecting a more compact set of features, the model not only delivers better results but also becomes more efficient and cost-effective; fewer inputs translate into lower computational cost and faster processing, which is practical for real-world use. SHAP analysis further validates the importance of the selected features—such as IDH1, EGFR, and NF1—and, importantly, makes each prediction explainable, so that clinicians can see why a case was flagged and discuss the reasoning in tumor-board or radiology review.

Beyond technical metrics, the intended clinical value is early and reliable support: high-confidence outputs with clear drivers can prompt faster biomarker confirmation and earlier escalation of complex cases, while low-risk outputs with diffuse drivers can proceed through routine pathways. To ensure that these benefits hold outside of development data, we will perform external validation on independent cohorts

from other institutions and time periods, calibrate predicted probabilities, and select operating thresholds with clinical input. We also plan a prospective pilot in routine workflow to measure impact on turnaround time, downstream testing, and patient management, while monitoring stability across seeds, splits, and patient subgroups. In this way, the system moves from strong performance on paper toward safe, useful, and sustainable deployment in practice, while maintaining efficiency and interpretability.

REFERENCES

[1] World Health Organization, "International Agency for Research on Cancer," https://gco.iarc.fr, n.d..

[2] Cleveland Clinic, "Glioma," https://my.clevelandclinic.org/health/diseases/21969-glioma, n.d..

[3] Miller, K.D., Siegel, R.L., Lin, C.C., Mariotto, A.B., Kramer, J.L., Rowland, J.H., Stein, K.D., Alvaro, M., Jemal, A., "Cancer statistics for adolescents and young adults, 2020," *CA: A Cancer Journal for Clinicians*, vol. 70, no. 6, pp. 443–459, 2020.

[4] Louis, D.N., Ohgaki, H., Wiestler, O.D., Cavenee, W.K., the WHO Classification of Tumours of the Central Nervous System, "The 2007 WHO classification of tumours of the central nervous system," *Acta Neuropathologica*, vol. 114, no. , pp. 97–109, 2007.

[5] Gaillard, F., "WHO Classification of CNS Tumors," https://radiopaedia.org/articles/who-classification-of-cns-tumours-1?lang=us, 2022.

[6] Hanif, F., Dastgir, G., Abbas, Q., Hussain, I., Ghaffar, S., "Glioblastoma multiforme: A review," *Asian Pacific Journal of Cancer Prevention*, vol. 18, no. 1, pp. 3–9, 2017.

[7] Mirchia, K., Richardson, T., "Beyond IDH-mutation in gliomas," *Cancers*, vol. 12, no. , pp. 1817, 2020.

[8] Vigneswaran, K., Esquenazi, Y., Melkonian, S.C., Chow, A., Chapman, M., Green, S., Min, B., Nasrallah, A., Lanchon, S., Pradhan, R., "Advances in glioma genetics," *Annals of Translational Medicine*, vol. 3, no. , pp. 95, 2015.

[9] DeWitt, J., Choi, M., Schwartz, A., Epstein, M., Jolly, M., Mishra, D., "Cost-effectiveness of IDH testing in gliomas," *Neuro-Oncology*, vol. 19, no. , pp. 1640–1650, 2017.

[10] Krauze, A., Manley, M., Jang, C., Olson, R., Saeed, R., Shukla, N., Das, S., "AI-Driven Image Analysis in CNS Tumors," *Journal of Biotechnology and Biomedicine*, vol. 5, no. , pp. 1–19, 2022.

[11] Diaz Rosario, M., Tohme, R., Islam, N., "Sex Differences in Glioma Data," *Biomolecules*, vol. 12, no. , pp. 1203, 2022.

[12] Liu, X., Zhang, L., Zhang, S., Wu, X., "Deep Learning in Glioma Prognosis: A Survey," *Frontiers in Oncology*, vol. 10, no. , pp. 798, 2020.

[13] Ko, C., Brody, J. P., "A genetic risk score for glioblastoma multiforme based on copy number variations," *Cancer Treatment Research Communications*, vol. 27, no. , pp. 100352, 2021.

[14] Sánchez-Marqués, R., García, V., Sánchez, J. S., "A data-centric machine learning approach to improve prediction of glioma grades using low-imbalance TCGA data," *Scientific Reports*, vol. 14, no. , pp. 17195, 2024.

[15] Kha, Y.-H., others, "Development and validation of an efficient MRI radiomics signature for improving the predictive performance of 1p/19q codeletion in lower-grade gliomas," *Cancers (Basel)*, vol. 13, no. 21, pp. 5398, 2021.

[16] Xia, X., others, "Interpretable Machine Learning Models for differentiating glioblastoma from solitary brain metastasis using radiomics," *Academic Radiology*, vol. , no. , pp. , to be published, 2025.

[17] Tang, Y., others, "Identification of five important genes to predict glioblastoma subtypes," *Neuro-Oncology Advances*, vol. 3, no. 1, pp. , 2021.

[18] Kumar, D., others, "An improved machine-learning approach for COVID-19 prediction using Harris Hawks Optimization and feature analysis using SHAP," *Diagnostics*, vol. 12, no. 5, pp. 1023, 2022.

[19] Nalasari, L. T., Anam, S., Shofianah, N., "Liver cirrhosis classification using extreme gradient boosting classifier and Harris Hawks optimization as hyperparameter tuning," *Journal of Electronics, Electromedical Engineering and Medical Informatics*, vol. 7, no. 2, pp. , 2025.

[20] Kurdi, S. Z., others, "Brain tumor classification using meta-heuristic optimized convolutional neural networks," *Journal of Personalized Medicine*, vol. 13, no. 2, pp. 181, 2023.

[21] Pirgazi, J., others, "An efficient hybrid filter-wrapper method based on improved Harris Hawks optimization for feature selection," *BioImpacts*, vol. , no. , pp. , 2024.

[22] Louis, D.N., Perry, A., Wesseling, P., Brat, D.J., Cree, I.A., Figarella-Branger, D., Hawkins, C., Ng, H.K., Pfister, S.M., Reifenberger, G., Soffietti, R., von Deimling, A., Ellison, D.W., "The 2021 WHO classification of tumours of the central nervous system," *Neuro-Oncology*, vol. 23, no. 8, pp. 1231–1251, 2021.

Federated Performance-based Averaging (FedPA): A Robust and Selective Learning Framework for Chest X-Ray Classification in Heterogeneous Data Environments

Atif Mahmood[1], Tashin Khan Sadique[2], Saaidal Razalli Azzuhri[3], Roziana Ramli[4], Leila Ismail[5]
Faculty of Data Science and Information Technology,
INTI International University, Nilai, Malaysia, 71800[1,2]
Department of Computer System and Technology-Faculty of Computer Science and Information Technology,
Universiti Malaya, Kuala Lumpur, Malaysia, 50603[3]
Department of Computer and Information Sciences, Northumbria University, Newcastle, UK[4]
Department of Computer Science and Software Engineering, United Arab Emirates University, UAE[5]

***Abstract*—Chest X-ray imaging remains a cornerstone in the diagnosis of thoracic conditions such as COVID-19, pneumonia, and lung opacity. Despite advancements in deep learning, the development of robust and generalizable models is limited by data privacy constraints, as patient data cannot be centralized across institutions. Federated Learning (FL) has emerged as a promising solution by enabling collaborative model training without sharing raw data. However, standard FL algorithms like FedAvg, FedProx, and FedSGD aggregate all client updates without considering their individual quality, making them vulnerable to performance degradation in the presence of data heterogeneity, label noise, or underperforming clients. To address these challenges, this study proposes Federated Performance-Based Averaging (FedPA), a novel selective aggregation strategy that incorporates only those client models that meet a predefined performance threshold during training. By leveraging an accuracy-based filtering mechanism, FedPA ensures that only sufficiently trained and reliable local models contribute to global updates. The method was evaluated on a multi-class, non-IID chest X-ray dataset containing four classes: Normal, COVID-19, Pneumonia, and Lung Opacity. Using DenseNet as the backbone model, experiments were conducted across four federated clients, each biased toward a specific class to simulate real-world data distributions. Results demonstrate that FedPA significantly outperforms baseline federated algorithms across key metrics, achieving a global accuracy of 91.82%, F1-score of 92.48%, and recall of 92.08%. The method also achieved faster convergence, higher stability, and reduced round-to-round accuracy fluctuations. System-level evaluations further show that FedPA offers competitive efficiency in terms of inference time, throughput, CPU usage, and memory footprint, making it suitable for deployment in resource-constrained clinical environments. Overall, FedPA offers a practical and effective advancement in federated learning for medical imaging. By filtering unreliable client contributions, it preserves model quality and privacy, presenting a viable path for clinical deployment in scenarios where data centralization is infeasible due to ethical, legal, or logistical constraints.**

***Keywords*—Public health; industrial growth; federated learning; FedAvg; FedPA; FedSGD**

I. Introduction

Deep learning [1] and Medical imaging has become an necessary tool in current healthcare, playing an important role in the diagnosis, monitoring, and treatment of various diseases [2]. Technologies such as X-rays, computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound, combined with AI-driven analysis (see Fig. 1), offer detailed insights into the human body, enhancing the ability of healthcare professionals to detect abnormalities at an early stage with greater accuracy and efficiency [3]. With rapid advancements in artificial intelligence (AI) and deep learning, medical image analysis has significantly improved [4], offering enhanced accuracy, speed, and efficiency in diagnosing conditions such as cancer [5], pneumonia [6], neurological disorders [7], and cardiovascular diseases [8]. These AI-driven approaches facilitate automated detection, segmentation, and classification of medical images, reducing the burden on radiologists and improving patient outcomes [9], [10].

The collection and management of medical imaging datasets for AI applications face significant challenges due to strict privacy regulations and ethical concerns. Federated Learning (FL) [11] has emerged as an innovative, privacy-preserving solution that enables collaborative model training across healthcare institutions without sharing raw patient data (see Fig. 1). By training models locally and only sharing model updates, FL ensures data privacy while leveraging distributed datasets [12]. Federated averaging (FedAvg) is a common aggregation procedure in federated contexts. FedAvg, however, experiences convergence issues, especially when there is significant diversity in the data distributions among clients [13].

In the next section, we will examine the role of federated learning in healthcare, highlighting its current applications and limitations. This discussion will also identify the existing research gaps, thereby framing the specific aim and contribution of this paper in addressing those challenges.

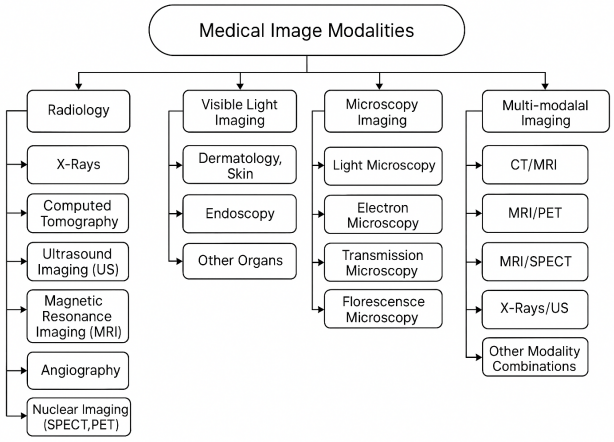


Fig. 1. Medical imaging modalities classification [3].

A. Federated Learning in Healthcare

In a standard FL setup, each client trains a local model using its own dataset and transmits model updates to a central server, which aggregates these updates to improve the global model. This iterative process continues until the model reaches convergence (see Fig. 2). However, traditional FL algorithms such as FedAvg struggle with non-IID data distributions, client heterogeneity, and communication overhead, particularly in healthcare applications where data availability and quality vary significantly across institutions.

Federated Learning has gained significant traction in the healthcare domain due to its ability to enable collaborative model training while ensuring compliance with data privacy regulations such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation) [14]. In healthcare, FL is particularly useful for applications such as medical imaging analysis, predictive diagnostics, personalized treatment planning, and disease progression monitoring.

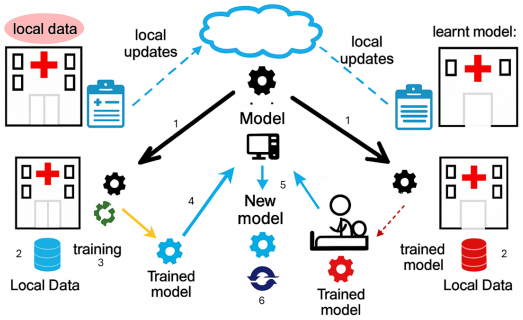


Fig. 2. Federated learning step by step process.

Recent studies have introduced various aggregation techniques to enhance federated learning, with Federated Averaging (FedAvg) being one of the most widely adopted due to its simplicity and efficiency [15]. However, FedAvg assumes equal participation from all clients, making it vulnerable to unreliable or delayed clients, particularly in real-world federated networks where client availability is inconsistent due to factors such as network failures, power constraints, and computational limitations. This inconsistency can lead to several challenges, including unstable global updates, where a high dropout rate among clients causes model updates to become erratic and degrade overall performance; inefficient optimization, as the server lacks control over client participation, making it difficult to streamline the learning process; and reduced model quality, since the absence of high-performing clients in certain rounds may lead to lower-quality updates.

To address some of these limitations, several advanced aggregation techniques have been proposed. Federated Learning with Proximal Optimization (FedProx) [16] improves FedAvg by mitigating the negative effects of client data heterogeneity. However, tuning its additional hyperparameter is challenging, and its default values may not generalize well across different datasets. Federated Multi-Task Learning via Multi-Task Association (FedMA) [17] optimizes both global and local models jointly, which enhances convergence, but its complexity makes it difficult to implement in some federated learning systems. Quantization-based Federated Learning (QFFL) [18] reduces communication overhead by compressing model updates through quantization, but this approach introduces quantization errors, which can impair convergence and final model accuracy.

Despite advancements in Federated Learning (FL), existing methods still face major challenges in medical imaging. High communication overhead, caused by frequent model updates, burdens low-bandwidth healthcare networks. Non-IID data due to variations in demographics, imaging tools, and protocols limits model generalizability.

Client participation is often inefficient, particularly from resource-constrained or outdated devices, contributing low-quality updates and hindering convergence. Data heterogeneity and noise further destabilize training. Scalability is also limited by diverse hardware capabilities across institutions.

Moreover, FL systems remain vulnerable to privacy and security threats such as model inversion and adversarial attacks. Unreliable client connectivity common in mobile or remote setups adds to participation inconsistency, complicating the learning process.

To address these limitations, we propose Performance-Based Federated Averaging (FedPA). A novel aggregation technique specifically designed for medical imaging applications. Unlike the traditional FedAvg method, which aggregates updates from all participating clients regardless of their quality, FedPA selectively incorporates only those updates that meet a predefined accuracy threshold (e.g., 80%). This selective inclusion significantly reduces noise from poorly performing clients, accelerates convergence, and enhances overall model robustness.

In addition to performance-aware selection, FedPA allows low-performing clients to download the updated global model parameters, enabling them to improve their local models over time. Once their performance meets the participation threshold, they are reintegrated into the aggregation process. This strategy promotes progressive learning and inclusivity while maintaining model quality and minimizing communication overhead.

To evaluate the effectiveness of FedPA, we applied it to a multi-class chest X-ray dataset for thoracic disease detection (Normal, COVID-19, Lung Opacity, Pneumonia). Each client was assigned a specific class, simulating a realistic non-IID setting. Experimental results demonstrate that FedPA outperforms traditional FL methods in terms of accuracy, F1-score, robustness, and communication efficiency highlighting its potential as a performance-aware and scalable FL approach for healthcare AI applications

The remainder of this paper is structured as follows. Section 2 presents a literature review of existing FL techniques in medical imaging, discussing the limitations of traditional aggregation methods. Section 3 details the methodology, including an overview of FL, the architecture of the proposed FedPA framework, model configurations, and dataset preprocessing. Section 4 provides experimental results and discussion, comparing FedAvg, traditional FedPA, and Performance-Based FedPA in terms of accuracy, loss, convergence time, and efficiency. Section 5 discusses the broader implications of the findings, highlighting advantages, challenges, and potential real-world applications of FedPA. Section 6 concludes the paper with a summary of findings and future research directions, focusing on further optimizations and the integration of secure aggregation techniques for enhanced privacy in FL.

II. Aggregation Strategies in Federated Learning

1) Federated Averaging (FedAvg): Federated Averaging (FedAvg) is the most commonly used FL aggregation strategy, introduced by [11]. In FedAvg, each participating client trains a local model for several epochs and sends the updated model weights to the central server. The global model is updated using the weighted average of all client models, formulated as:

$$w_{t+1} = \sum_{k=1}^{K} \frac{n_k}{N} w_k^t \tag{1}$$

where w_k^t is the local model update from client k, n_k is the number of samples on client k, and $N = \sum_{k=1}^{K} n_k$ represents the total number of data samples across all clients.

While FedAvg is effective in IID (independent and identically distributed) settings, it faces challenges in heterogeneous and non-IID environments, where certain clients may contribute low-quality updates, slowing down convergence.

Federated learning studies in medical imaging commonly use FedAvg as the standard baseline, often comparing it with optimization-enhanced and personalized alternatives. For instance, [19] report an accuracy approximately 97% for FedAvg under IID conditions, while personalized approaches in [20] and [21] show AUROC values of 0.95 for FedAvg versus 0.94 for local models. These studies cover diverse imaging modalities such as MRI, CT, X-ray, OCT and dataset sizes ranging from 2,100 to over 84,000 images. Methods tailored to handle data heterogeneity, such as FedOpt, often demonstrate improved accuracy and robustness under non-IID conditions. However, most of these approaches (six out of seven) incur higher client-side computational costs, and communication overhead typically remains moderate to high.

The [22] study shows that using EfficientNet-B0 with the FedAvg algorithm in a federated learning framework improves privacy and diagnostic accuracy for MRI brain tumor detection. EfficientNet-B0 outperforms ResNet in handling data heterogeneity, emphasizing the potential of federated learning for robust medical image analysis.

Referred by [23] SecureFed, a secure federated learning aggregation method, outperforms FedAvg, FedMGDA+, and FedRAD in lung abnormality analysis using chest X-rays. It demonstrates superior robustness and fairness on COVID-19 datasets and is adaptable for multimodal medical data analysis.

The [24] study introduces Auto-FedAvg, a data-driven federated learning approach that dynamically adjusts aggregation weights based on data distribution and training progress. It outperforms existing methods on heterogeneous CIFAR-10 and proves effective in two medical imaging tasks: COVID-19 lesion segmentation in chest CT and pancreas segmentation in abdominal CT.

While FedAvg is simple and effective, it suffers from inefficiencies when dealing with non-IID data and heterogeneous client performance, leading to slower convergence and suboptimal accuracy.

2) Federated Proximal (FedProx): To address the limitations of FedAvg, FedProx [16] introduced a proximal term to stabilize local updates, especially when clients have heterogeneous computing power or varying data distributions. The modified objective function is:

$$w_k^{t+1} = \arg\min_w \left(F_k(w) + \frac{\mu}{2} \|w - w_t\|^2 \right)$$

where u controls how much a client update deviates from the global model. FedProx helps mitigate drift in non-IID settings but does not directly address client selection based on performance.

The author in [25] presents a comparative analysis of federated aggregation algorithms for binary classification of X-ray images across a limited number of hospitals with varying data heterogeneity. Among the evaluated methods, FedProx consistently outperforms others, making it the most effective approach in handling statistical variation in distributed medical imaging settings.

Referred by [26], FedProx outperforms FedAvg in federated learning for medical image analysis, demonstrating superior performance in cancer classification despite heterogeneous client data. However, increasing the number of communication rounds between the server and clients degrades model performance, affecting convergence and accuracy.

The author in [27] introduces a hybrid federated learning algorithm, FedProx:FedSplit, designed to tackle both statistical and system heterogeneity in medical data communication. This approach enhances model convergence and improves the overall accuracy of the global model in distributed healthcare settings.

This [28] study explores brain tumor detection using MRI images in a federated learning (FL) setup to preserve data privacy. A VGG19-based model was trained across four clients with non-IID data, incorporating Grad-CAM for explainability. Using a centralized test set for fair evaluation, the model achieved high accuracy of 97.18% using FedAvg, 98.24% using FedProx, and 98.45% using Scaffold. It demonstrating the effectiveness of FL in privacy-preserving, real-world medical imaging scenarios.

3) Federated Proximal (FedBN): Federated Batch Normalization (FedBN) [29]is a variant of federated learning designed to tackle non-IID (non-identically distributed) data across clients by adjusting how batch normalization layers are handled. In standard federated learning (e.g. FedAvg), all model parameters (including batch normalization statistics) are averaged across clients, which can be problematic when clients have very different feature distributions (a situation known as feature shift). FedBN addresses this by keeping the batch normalization parameters local to each client. In other words it does not include the BN layers running mean and variance in the global model aggregation [30].

Medical imaging federated learning often faces site-specific data distributions for example, MRI scans from different hospitals may have varying intensity distributions due to different scanners or protocols, even if the underlying task (e.g. tumor classification) is the same [29]. FedBN is well-suited to such settings: by not forcing a single global normalization, it allows each hospitals model to adjust to its own imaging characteristics while still contributing to a shared global model. This approach has been applied in a number of medical imaging tasks to improve generalization across institutions. For instance, researchers have evaluated FedBN on multi-center datasets for diagnostic classification and segmentation problems, observing better performance compared to vanilla federated training when data distributions differ [31]

Notably, FedBN has been used as a state-of-the-art baseline in: Histopathology Image Classification Cell Nuclei Segmentation MRI Segmentation The method's success in these peer-reviewed studies demonstrates how personalized normalization can boost performance in privacy-preserving medical AI collaborations.

While FedBN was introduced to address the issue of batch normalization (BN) layers in a non-IID setting. In standard BN, statistics are computed from local mini-batches that might not generalize properly across clients. FedBN sidesteps BN layers during aggregation, allowing the client-specific normalization parameters while aggregating the remaining ones. This modification improves generalization, thus preventing the collapse in performance due to misaligned feature statistics [13].

In FedBN, the model parameters are separated into two components:

$$w = \{w_{\text{shared}}, w_{\text{BN}}\}$$

where:

- w_{shared} are the weights of shared layers (e.g., convolutional layers),
- $w_{\text{BN}} = \{\gamma, \beta, \mu, \sigma^2\}$ are the batch normalization parameters: scale, shift, running mean, and running variance.

The FedBN update rules are then:

$$w_{\text{shared}}^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{n} w_{\text{shared},k}^{(t)} \tag{2}$$

$$w_{\text{BN},k}^{(t+1)} = w_{\text{BN},k}^{(t)} \quad \text{(not aggregated)} \tag{3}$$

This means batch normalization parameters remain local to each client to better handle feature distribution heterogeneity.

A. Federated Stochastic Gradient Descent (FedSGD)

FedSGD is a foundational algorithm in federated learning that updates the global model by aggregating gradients from multiple clients rather than full model weights. Unlike FedAvg, which averages model parameters after multiple local updates, FedSGD performs only one gradient computation per client per round, reducing local computation but increasing communication cost. The global objective is to minimize the weighted empirical risk:

$$\min_{\theta} f(\theta) = \sum_{k=1}^{K} \frac{n_k}{n} f_k(\theta)$$

Each client k computes its local gradient using the current global model θ^t:

$$g_k^t = \nabla f_k(\theta^t)$$

The server aggregates gradients from the selected clients $\mathcal{S}_t$ and updates the global model as:

$$\theta^{t+1} = \theta^t - \eta \sum_{k \in \mathcal{S}_t} \frac{n_k}{n_{\mathcal{S}_t}} g_k^t$$

If uniform weighting is used instead of sample-size-based weighting:

$$\theta^{t+1} = \theta^t - \eta \cdot \frac{1}{|\mathcal{S}_t|} \sum_{k \in \mathcal{S}_t} g_k^t$$

This approach ensures lightweight local computation and is suitable for bandwidth-constrained distributed settings.

FedAc [32] can achieve a linear speedup with fewer rounds of synchronization compared to FedAvg, improving communication efficiency. FedAc has stronger theoretical guarantees than FedAvg, particularly for functions that are third-order smooth. The authors developed a novel analysis approach and a strategic tradeoff between acceleration and stability to derive the FedAc algorithm.

The key idea of the FedSel framework [33] is to privately select the top-k most important dimensions in each iteration of

federated SGD, in order to reduce the noise injected into the gradients under local differential privacy. The authors claim that their FedSel framework outperforms the state-of-the-art solutions in terms of privacy, accuracy, and time complexity.Experiments on real-world and synthetic datasets verify the effectiveness and efficiency of the FedSel framework.

The paper [34] analyzes the convergence of local descent methods like Federated Stochastic Gradient Descent (FedSGD) for solving nonconvex optimization problems in federated learning with heterogeneous data.

Despite notable advances in federated aggregation strategies such as FedAvg, FedProx, FedBN, and FedSGD, existing approaches remain limited in their ability to adaptively account for client performance quality under highly heterogeneous and non-IID medical data. Most methods either assume uniform participation (FedAvg), add complexity with sensitive hyperparameters (FedProx), or address only feature shift without considering contribution reliability (FedBN), while lightweight approaches like FedSGD incur high communication costs without performance-based filtering. Furthermore, recent personalized and optimization-driven extensions improve convergence but often impose higher computational and communication overhead, making them less practical for resource-constrained healthcare environments. A clear gap therefore exists in the literature for a performance-aware, lightweight, and selective aggregation mechanism that not only filters out low-quality updates but also ensures inclusivity for weaker clients to progressively improve an area where FedPA positions itself as a novel and practical contribution.

III. Federated Learning Performance-based Averaging (FedPA)

Federated Averaging (FedAvg) is one of the most widely adopted aggregation methods in federated learning due to its simplicity and communication efficiency [15]. However, it assumes uniform participation and contribution quality from all clients, which rarely holds in real-world settings. Client availability often varies due to network instability, power constraints, and hardware limitations, leading to unreliable or delayed updates. This inconsistency results in unstable global model updates, inefficient optimization, and reduced model quality especially when high-performing clients are absent from certain rounds. Furthermore, FedAvg lacks mechanisms to differentiate between high- and low-quality updates, allowing poor local models to degrade overall performance, particularly when data quality, size, or computational resources vary significantly across clients.

To overcome this limitation, Performance-Based FedPA is introduced. This approach integrates a client contribution evaluation mechanism that selectively aggregates model updates from high-performing clients while still allowing lower-performing clients to participate and improve. This strategy ensures that the global model prioritizes high-quality updates, leading to improved overall performance and stability.

Step 1: Initialization

- Initialize the global model w_0.
- Set the performance threshold τ (e.g., 80% accuracy).
- Define all participating clients K.

Step 2: Federated Learning Process

For each global round $t = 1, 2, \ldots, T$:

- Server selects a subset of clients $S_t \subseteq K$.
- Each client $k \in S_t$:
  - Trains its local model w_k^t using its dataset.
  - Evaluates its local model accuracy:

$$\text{Accuracy}(w_k^t)$$

  - If $\text{Accuracy}(w_k^t) \geq \tau$, the client is added to the selected clients set C_{selected}:

$$C_{\text{selected}} = \{k \mid \text{Accuracy}(w_k^t) \geq \tau\}$$

  - If $\text{Accuracy}(w_k^t) < \tau$:
    - The client still receives the global model update but does not contribute to aggregation.
    - This ensures continuous learning for weaker clients.

Step 3: Optimization and Weight Computation

For each selected client $k \in C_{\text{selected}}$:

- Compute the contribution weight α_k based on accuracy:

$$\alpha_k = \frac{\text{Accuracy}(w_k^t)}{\sum_{j \in C_{\text{selected}}} \text{Accuracy}(w_j^t)}$$

- Compute the local model update Δw_k^t:

$$\Delta w_k^t = w_k^t - w_t$$

Step 4: Aggregation Using FedAvg

Once all selected clients send their updates to the server, the global model is updated using FedAvg.

Using the weighted averaging formula:

$$w_{t+1} = \sum_{k \in C_{\text{selected}}} \frac{\alpha_k}{N} w_k^t$$

where:

- $N = |C_{\text{selected}}|$, the number of selected clients.
- Stable and effective for non-IID data.

- Reduces variance in updates.
- Slower convergence but better final accuracy.

Step 5: Updating the Global Model

- Server broadcasts the updated global model w_{t+1} to all clients.
- The process repeats for T rounds.
- At the end, the final model w_T is returned.

A. Advantages of Performance-Based FedPA

1) Improved global model accuracy: By prioritizing contributions from high-accuracy clients, the global model avoids the negative impact of poor updates, leading to faster and more stable convergence. Ensures that well-trained models influence learning while still allowing weaker models to improve.

2) Faster convergence: Since only high-performing models significantly influence the global update, the model learns more effectively per round, reducing the number of communication rounds required for convergence. Avoids instability caused by noisy updates from clients with poor local training.

3) Adaptive participation for weaker clients: Clients not meeting the accuracy threshold are still included in the federated learning process, receiving global updates and improving over multiple rounds. This encourages progressive learning, allowing weaker models to gradually enhance their accuracy and contribute more effectively in future rounds.

4) Robustness to Non-IID data: Real-world federated learning scenarios, such as multi-institutional medical imaging, often involve non-IID (heterogeneous) datasets. Performance-Based FedPA naturally adapts to these conditions by giving higher importance to clients with better model performance, which mitigates the impact of data heterogeneity.

IV. Methodology

A. Data Acquisition and Preprocessing

This study is based on dataset images of the chest radiograph classifying images into four classes-Normal, COVID-19, Viral Pneumonia, and Lung Opacity. Data from this collection used for the study is from publicly available clinical repositories to make sure the patient's demographics and radiological conditions are as diverse as needed. Such diversity will strengthen the generalization capability of the trained models for better and more effective application in real-world clinical conditions (Fig. 3).

Each image was preprocessed through a similar standard pipeline that was used across federated clients. All images were resized to a consistent input resolution of 224x224 pixels, which is widely adopted in convolutional neural network (CNN) models. Afterward, the pixel intensities were normalized by scaling the values into the range from [0, 1]. This definition of normalization was further adjusted per channel by mean and standard deviation values calculated on the training set to standardize input data of different images.

Various data augmentation techniques were utilized during training to reduce overfitting and improve the generalization abilities of the model. These augmentation methods included random horizontal flipping, small-angle rotations, random cropping, and intensity rescaling. The augmentations mimic image acquisition variability across the various institutions and imaging equipment, which is one of the main requirements for building robust models in medical image analysis. These techniques improve model robustness by preparing a model that performs better in real data differing from the acquisition conditions and quality.

All the changed images, label and metadata have been stored in an organized NumPy archive (covid_dataset.npz) to directly load and access such data during the training operation. This format is readily available for any federated client to train on such dataset providing a facility to have a smooth Read-Write process in distributed learning in healthcare AI systems.

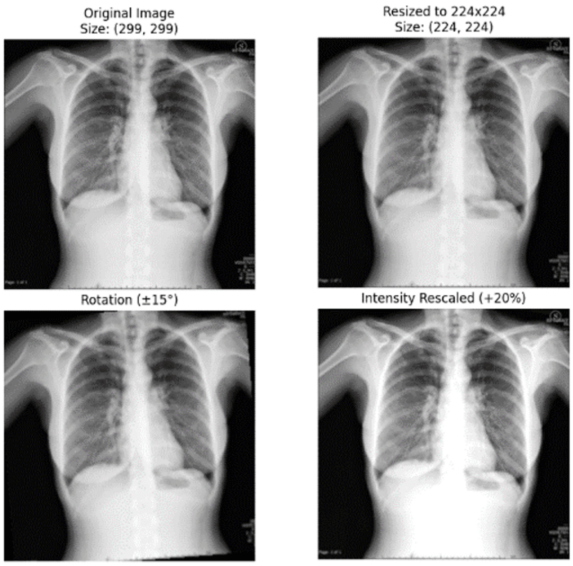


Fig. 3. This figure shows the preprocessing pipeline, the original image is resized into 224 x 224 size. This is further rotated and has its intensities increased for better feature recognition and generalization.

B. Data Partitioning and Distribution Strategy

The partitioning of the preprocessed chest X-ray dataset into four virtual clients simulating different hospitals or settings under federated learning, therefore representing three thoracic diseases-COVID-19, Pneumonia, and Lung Opacity-and one Normal class, would provide insight into these instances' diverse medical diagnosis as seen in real-world clinical practice. A truly non-IID data partitioning approach for federated learning algorithms is justified because, in practice, medical data can vary from one institution to another depending on specialization, as well as geographical attractiveness for certain diseases. Now, the data of each client was divided into a training set (70%), a validation set (15%), and a global test set (15%). The training and validation data were kept by the local class distributions, whereby the model of each client's training learned from an equal balance of images to prevent any biases

that could have arisen from unequal class proportions. The global test set, which holds 15% of the total dataset, was evenly distributed across all classes, allowing it to be completely separate to ensure an unbiased evaluation of the aggregated model's performance. This partitioning strategy exploits natural statistical variability that different clients have for the evaluation of federated learning methods under decentralized, imbalanced, and real-world clinical data conditions. Such a mechanism becomes necessary to allow for the generalization of the trained models over dissimilar clinical terrains, while addressing nature issues like data heterogeneity and class imbalance.

C. Model Architecture

Proposed model works on deep learning algorithms to classify chest X-ray images under four categories: Normal, COVID-19, Lung Opacity, and Pneumonia. The architecture promises extraction of multi-level spatial features with a reduction in computational complexity, thereby making it an ideal flaunt for federated learning settings.

1) The input layer: images of size 224 x 224 in grayscale feed into the model such that all the images could also undergo pre-processing in the same way for all federated clients.

2) Convolutional layer: Three convolutional blocks are required to capture the most relevant features such as edges, textures, patterns related to abnormalities in the lungs, and consist of a convolutional layer each followed by ReLU activation and max pooling.

3) Flatten layer: This layer converts 2-D feature maps made from the convolutional layers into single one-dimensional vectors for subsequent processing.

4) Dense layer: This is a dense layer built with 128 neurons and by ReLU activation to model very complex behaviours. A dropout layer is included at 30% rate to reduce overfitting and thus enhance generalization across clients.

5) Output layer: There are 4 neurons that correspond to classes using the SoftMax activation function for multi-class classification.

The model is optimized with the categorical cross-entropy loss for multi-class classification and is best suited for fast training by the Adam optimizer. It is a lightweight model that affords on-the-fly training on pretty cheap machines on the client-side in a federated learning system.

D. Global and Local Model Training in Federated Learning

In federated learning, model training involves two main phases: local model training and global model aggregation. These phases are iteratively repeated to make the global model better while ensuring data privacy.

1) Client model training: The local model for a specific client will be trained based on its local private data and the updates to be sent to the global server will only be that specific model updates such as weights or gradients. The training procedure involves:

a) Model initialization: Clients receive the global model weights at the beginning of the training round.

b) Local epochs: Clients train their model for a set number of epochs on local data.

c) Update sharing: After training, each client computes the updates (model weights or gradients) before sending them back to the global server.

This decentralized structure ensures that client sensitive data never leaves clients.

2) Global model training: The global server aggregates all local updates to improve the global model after all the local updates have been collected. This will include averaging the weights of all models from participating clients after the process. The most common aggregation algorithms include the following:

a) FedAvg: This is basically the first federated learning algorithm where the server computes a weighted factor average of model updates to carry out its aggregation. This would be especially useful when the distribution of data was relatively similar among the clients.

b) FedProx: This adds a proximal term in a local objective with respect to deviation from the global model for local models. This is usually good where very much dissimilar data exist across clients.

c) FedBN: This is a specific method designed for very different data distributions across client bases, as might occur in medical imaging tasks. FedBN keeps different batch normalization statistics for each client to handle domain shifts, thereby completely solving the problem of tasks.

d) FedSGD: Unlike FedAvg, which uses model weights, FedSGD is used to aggregate gradients instead of model parameters. It thus minimizes data exchange between clients and the server, thus optimally suited under low-latency communication networks.

e) FedPA (Federated Performance Averaging): With this new method, performance-based client selection is introduced. So in FedPA, a client must surpass a certain performance threshold to add its model updates to the global model.

Ensure that the federated learning algorithms harness the global model through the knowledge gained from each client while having local data private. They are especially concerned with healthcare and medical imaging since they leverage data diversity to deal with privacy alongside efficient model aggregation (Fig. 4).

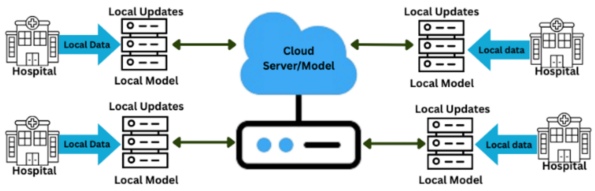


Fig. 4. Federated learning environment whereby hospitals learn and update local models on local patient data and use these to periodically update a shared central server for collaborative model updates without patient-data sharing.

V. EXPERIMENTS

A. Hardware and Software Requirements

The experiments were conducted on a workstation running Windows 11, equipped with an Intel Core i7 processor, 16 GB of RAM, and an NVIDIA RTX 4050 GPU with 12 GB of VRAM. The software environment was configured with Python 3.10, PyTorch 2.0.1, Torchvision 0.15.2, NumPy 1.24, scikit-learn 1.2, tqdm 4.65, and the CUDA Toolkit version 11.7. Model training was accelerated using CUDA where applicable, leveraging the available GPU resources. All models were implemented in PyTorch and executed in a single-machine, single-GP

B. Hyperparameter Settings

A careful selection of hyperparameters is critical in federated learning, as it significantly influences convergence speed, model generalization, and robustness to non-IID medical data distributions. This aligns with prior research in federated medical imaging and deep learning optimization. To ensure fair evaluation across all methods, the hyperparameters were standardized, drawing on empirical validation and insights from earlier studies addressing model regularization, optimizer behavior, and communication efficiency in federated settings.

Throughout all experiments, consistent hyperparameter settings were maintained. The AdamW optimizer was chosen for its improved handling of weight decay over standard Adam, with a stable initial learning rate of 0.001. A cosine annealing learning rate scheduler was applied at every epoch to adjust the learning rate dynamically. A weight decay of 0.01 was used to mitigate overfitting. The loss function employed was cross-entropy loss, a standard choice for classification tasks. A batch size of 16 was selected to balance memory usage and gradient estimation accuracy. Each client performed three local epochs per round, and all federated learning variants (FedAvg, FedProx, FedBN, and FedPA) were trained over 14 communication rounds. Data augmentation techniques, including random horizontal flips and rotations, were applied during training to enhance generalization. A DenseNet-based convolutional neural network served as the common base architecture across all approaches.

These hyperparameter values were fixed for all experiments to ensure both comparability and reproducibility, in line with best practices recommended in recent federated learning studies focused on medical imaging.

C. Client Configuration and Performance-Based Participation

As viewed from this research perspective, the federated learning setup imitates the real-life medical setting where the data distributions are inherently non-IID and client-specific. Four clients were set up to represent the four separate clinical categories from the chest X-ray dataset: Normal, COVID-19, Lung Opacity, and Pneumonia. Each of the four clients had access exclusively to its own set of labeled images, creating heterogeneity in class distribution and simulating isolated data silos of hospitals or diagnostic centers.

In the training process, each client was intended to keep optimizing its copy of the global model independently for five epochs per round using local data. This practice seems to fit well with reality since in federated health care, local computation is generally chosen over frequent communication due to privacy and latency issues. Nevertheless, global performance could be compromised if all clients participated uniformly, as when some clients perform really poorly due to poor data quality or too few training samples.

Consequently, the proposed FedPA framework implemented a performance-based client selection mechanism. At the end of every round, the clients assessed their local model(s) on a validation split and saved the best checkpoint. Only clients that achieved at least 85% validation accuracy were eligible to take part in the global aggregation. Generalization and robustness in federated learning can benefit from contribution-based client selection, as earlier studies have pointed out..

This participation filter remained in effect throughout the 15 communication rounds. FedPA, by removing low-quality model updates, fosters stable learning and reduces the risk of drift in the central model, which is often due to unreliable model updates by clients; a challenge posed in federated scenarios by heterogeneous and imbalanced datasets.

D. Global Aggregation Strategy

For every modification of the global model, there was an aggregation of local models by the participating clients at the end of each communication round, though significantly different regarding the mechanism of aggregation. In FedAvg, aggregation used to be done by the average of all client updates uniformly, presuming the reliability of all participants is equal. On a somewhat similar note, FedProx did not introduce any method of aggregation but provided a proximal term to recursively impose on local training so that students would penalize divergence from the global model within the locality, indirectly affecting aggregation quality. For another communication-intensive approach, FedSGD used gradients from each client, not the model updates but only required synchronous updates and higher communication overhead. FedBN performed normal averages on parameters but excluded all batch normalization layers from the averaging process such that the client held domain-specific statistics on normalization-an important feature in non-IID settings.

By contrast, the proposed FedPA approach redefined aggregation through a selection mechanism based on performance: It's not the fact that all clients contributed updates that matters, but instead, FedPA had just included clients whose local model outperformed a given validation accuracy threshold-all the way in this case, at or above the bar of 85 percent. Clients not meeting this criterion are not included in that specific round's aggregation. The selective strategy ensured that noisy and inferior updates were filtered out and only the well-trained local model participated in the global parameters. Thus, FedPA offered a greater quality enhancement under client performance and data heterogeneity-influencing factors concerning simple convergence and robustness, primarily in complicated evidence-dependent contexts, for example, medical imaging..

E. Evaluation Metrics

The performance of the federated learning models was evaluated based on the following main metrics

- Accuracy: It is the metric that indicates the overall performance of the model at classifying chest X-ray images correctly.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (4)$$

- Precision: Indicates the ability of the model to reduce false positives by correctly identifying only true positive instances.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

- Recall : Reflects the model's ability to correctly identify all positive instances, minimizing false negatives.

$$Sensitivity = \frac{TP}{FN + TP} \quad (6)$$

- F1 Score: It combines precision and recall into one measure and therefore sets a balance between the two by calculating their harmonic mean.

$$F1\ score = 2 * \frac{precision * recall}{precision + recall}. \quad (7)$$

- Inference time (s): Time needed to make predictions based on an input offered to the model by the user Additional measurements, such as CPU usage, memory usage, average inference time ,total inference time, throughput , were examined to determine the computational efficiency of the models.

VI. Result and Analysis

This section portrays a comparative analysis of the five federated learning models such as FedAvg, FedBN, FedPA, FedProx, and FedSGD. Evaluated on a common test set under exactly identical hardware and software conditions. The effect of the models is analyzed in determining the best among them in terms of the trade-off between performance in classification and efficiency at the system level for real-world deployment in almost all medical applications, such as chest X-ray analysis (Table I).

TABLE I. Performance Metrics Comparison

Method	Accuracy	Precision	Recall	F1 Score
FedAvg	0.9140	0.9259	0.9157	0.9206
FedBN	0.9124	0.9326	0.9053	0.9180
FedPA	0.9182	0.9295	0.9208	0.9248
FedProx	0.9113	0.9295	0.9071	0.9175
FedSGD	0.9129	0.9296	0.9102	0.9192

The standard evaluation metrics chosen for models are accuracy, precision, recall, and F1 score; each of these has its own clinical implications, especially in tasks related to disease detection.

1) Accuracy: FedPA obtained an accuracy of 91.82%, which is greater than FedAvg with 91.39% and FedSGD with 91.29%. This was achieved by performing client selection based on performance, excluding undertrained or noisy updates, which tended to minimize global model variance and aid generalization. FedAvg and FedSGD combine all clients equally, thus muting the effect of high-performing clients and slowing down convergence.

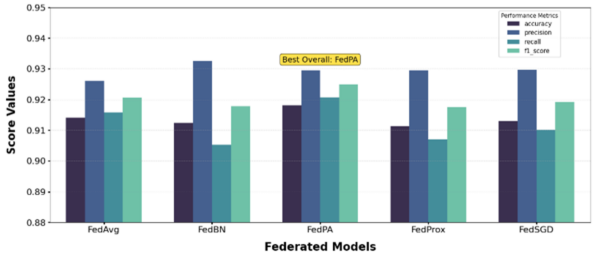


Fig. 5. Comparison of predictive performance metrics (Accuracy, Precision, Recall, F1 Score) of various federated models.

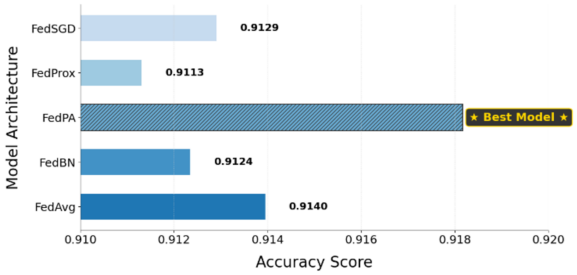


Fig. 6. Comparison of Accuracy across multiple federated models.

Precision FedBN leads in precision at 93.26% due to the use of client-specific batch normalization, thereby helping the model adapt to non-IID data distributions across clients. FedPA, close behind at 92.95%, strikes a good balance in precision without sacrificing too much recall. The implication is that FedBN really tries not to miss negative samples while FedPA has a better clinical trade-off by reducing the risk of under-detecting actual cases.

2) Memory: With 92.08%, FedPA achieves the highest recall performance, which indicates that it captures a lot of true positives, especially in lives-critical contexts like conversations about the identification of COVID-19 or pneumonia. This selective aggregation from FedPA contributes to this recall because high-performers among clients amplify the visibility of their presence to the global model and, hence, facilitate better identification of minority or diffused class patterns, even in the presence of data heterogeneity.

F1-ScoreFedPA has again captured the F1 score of 92.48%, thereby making it possible for it to prove its efficiency with respect to false positive or false negative prediction. With FedSGD and FedAvg coiled very closely behind, neither has the robustness of FedPA because it does not filter clients. The harmonic property of the F1 score underlines how selective aggregation according to FedPA effects compromise between sensitivity and specificity, both very important in terms of clinical reliability (see Fig. 5 and 6).

A confusion matrix is a table used to evaluate the performance of a classification model by comparing predicted labels with true labels. It is especially useful for understanding the types of errors a model makes. Here's the general structure for a binary classification confusion matrix (see Fig. 7 to 11):

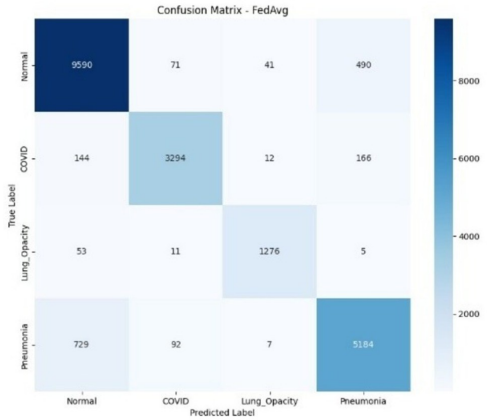


Fig. 7. Confusion Matrix of FedAvg.

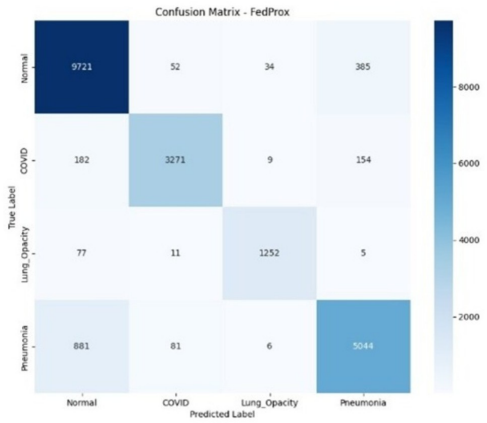


Fig. 8. Confusion matrix of FedProx.

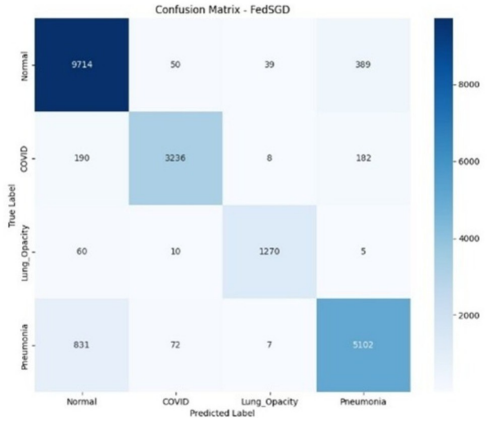


Fig. 9. Confusion matrix of FedSGD.

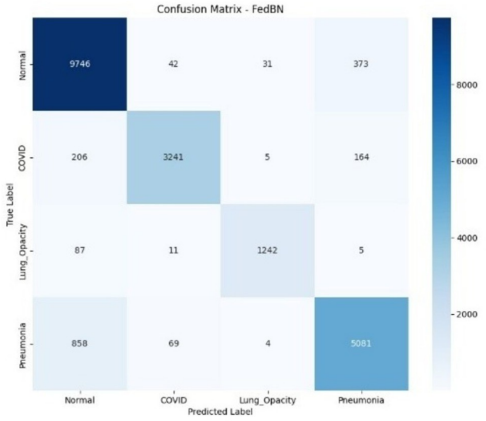


Fig. 10. Confusion matrix of FedBN.

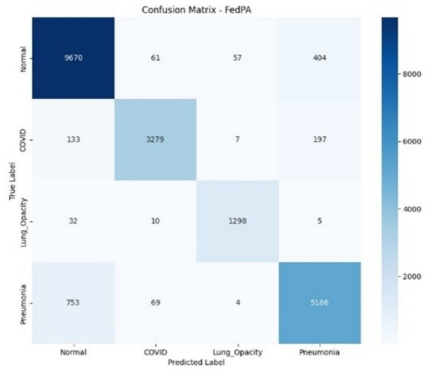


Fig. 11. Confusion matrix of FedPA: Proposed method.

A. System-Level Evaluation Metrics

This is now an assessment by which each model is compared for efficiency mainly due to the inference time, throughput, CPU usage, and memory footprint that are necessary for real-time medical AI deployment (Table II).

TABLE II. System Metrics Comparison

Method	Average Inference Time (ms)	Throughput	CPU Usage (%)	Memory Usage (MB)
FedAvg	0.003427	9328	13.15	12976
FedBN	0.002838	11266	13.11	13013
FedPA	0.002810	11377	12.85	13028
FedProx	0.002794	11442	12.98	13040
FedSGD	0.002812	11368	12.13	13053

1) Inference time: In terms of inference time, FedProx and FedPA are the fastest (2.8 ms/image), achieved by lightweight parameter updates and subsequently reduced model divergence during training. FedAvg comes in at 3.4 ms, likely due to model instability because of noisy client aggregation.

2) Throughput: FedProx attains the highest throughput (11,442 images/s) closely followed by FedPA (11,376) and FedSGD (11,368). These models benefit from efficient forward pass computation; besides, FedPA benefits from stabilized model updates that prevent computational spikes during testing. FedAvg, with a throughput of 9338, is thus lower due to its slow inference time and an increase in computational load.

3) CPU usage: FedSGD is the lightest on the CPU (12.13%), followed by FedPA (12.85%). Their minimal effort on steep client-side operations and sparse updates reduces compute requirements, thus lending themselves to edge environments. The higher CPU usage of FedAvg (13.15%) is likely due to its dense parameter averaging additionally worsened by client drift in non-IID settings.

4) Memory usage: Models cluster around 13 GB, with FedSGD (13.05 GB) and FedPA (13.11 GB) being somewhat lighter on memory. Since the overhead of memory and CPU power is manageable, these systems may also be deployed on a resource-limited scenario like mobile clinics (Fig. 12).

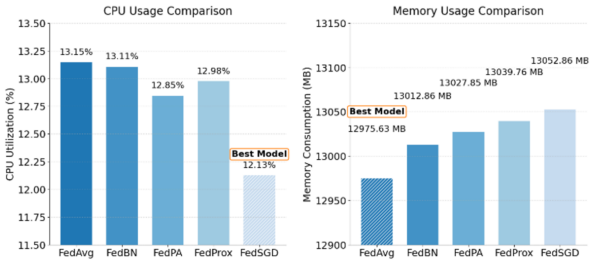


Fig. 12. Relative comparison of system resource usage between models with CPU usage and memory usage set side by side to emphasize the computational efficiency of the models.

B. Summarized Key Insights

We combine performance measures and systems metrics to identify the best model, lightest model, and most balanced model.

1) Best performing model: FedPA is characterized by high accuracy, recall, and F1 scores with low inference time and moderate resource consumption. Its inclusion strategy includes only reliable updates in the global model computation, thereby increasing the robustness of the FedPA in non-IID settings.

2) Most lightweight model: The FedSGD is by far the most lightweight in terms of CPU and memory resource consumption while still keeping throughput at a high level. It does have slightly less accuracy in the classification realm, but it is best suited for scenarios where resource constraints apply.

3) Most balanced model: FedPA, once again, sets the mark by balancing state-of-the-art accuracy with reasonable performance metrics. FedPA considers both efficiency and accuracy in diagnosis, a crucial capability for healthcare AI. In contrast, FedProx sacrifices accuracy for throughput (Fig. 13).

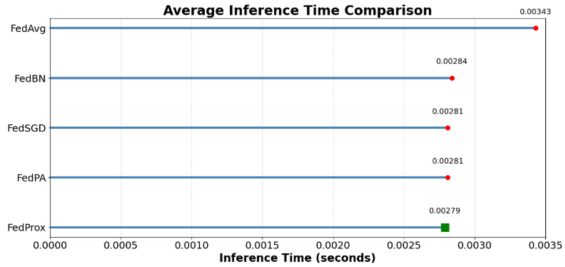


Fig. 13. Spider chart comparison of all eight test measures across all the models with a focus on the general predominance of FedPA regarding predictive performance, system efficiency, and consumption of resources.

C. Strengths of FedPA in Federated Learning

It is performance-aware aggregation that FedPA relies on for its superiority, where applying the client changes under a local accuracy condition of 85% would filter out updates and speed convergence, all while not falling into the trap of overfitting for the dominant client distribution.

Unlike FedAvg and FedProx, which accept all updates equally, FedPA selectively amplifies clients not by each, but only by those contributing meaningfully to learning and thus reducing negative transfer-the well-known problem of non-IID federated settings. It is both economical of communication bandwidth use, and doesn't waste computation cycles on low-quality updates; it is hence scalable and robust.

High recall and F1-score metrics further justify the clinical relevance of FedPA, especially in contexts where increased false negatives can lead to fatalities. Moreover, it is extremely stable across rounds and generalizes strongly; hence, it is earmarked for federated medical applications in which the parameters of accuracy and reliability are critical.

VII. Conclusion

This study introduced Federated Performance-Based Averaging (FedPA), a performance-aware federated learning approach designed to address challenges of data heterogeneity, variable client performance, and privacy concerns in medical imaging. Unlike traditional aggregation strategies such as FedAvg, FedProx, and FedSGD, which combine all client updates indiscriminately. FedPA selectively incorporates only high-performing client models based on a predefined accuracy threshold. This strategy helps to mitigate the negative impact of undertrained or noisy clients on global model performance.

In the context of COVID-19 chest X-ray classification, FedPA consistently outperformed other federated algorithms across all major evaluation metrics, achieving the highest accuracy, recall, and F1-score. It also demonstrated faster convergence and greater training stability, with reduced performance fluctuations and better retention across both strong and weak clients.

Looking ahead, several enhancements can further strengthen FedPA. These include adaptive thresholding that dynamically adjusts performance criteria across training rounds, and techniques to handle label noise through confidence-based filtering. Incorporating federated hyperparameter optimization methods may improve

convergence and fairness, while expanding to multi-modal data such as CT, MRI, and electronic health records could broaden the model's clinical utility. Real-world validation across geographically diverse hospital networks would also be essential to assess FedPA's scalability, generalization, and equity in healthcare delivery.

Overall, FedPA represents a significant advancement for privacy-preserving federated learning in healthcare. Its strong diagnostic performance, stability, and system efficiency make it a practical candidate for deployment in real-world clinical environments where centralized data access is limited.

A. Future Work

In this study, the evaluation of the proposed FedPA framework primarily focused on accuracy as the performance threshold for client selection. While this provided promising improvements in robustness and convergence, accuracy alone may not sufficiently capture the impact of class imbalance, which is a common challenge in medical imaging tasks. As part of future work, we aim to extend the thresholding mechanism by incorporating the F1-score alongside accuracy, thereby ensuring that both precision and recall are adequately considered in performance-based aggregation. This dual-metric approach will allow the framework to better handle skewed class distributions and minority class detection. Additionally, we plan to design adaptive selection strategies that dynamically choose the most suitable performance metric (e.g., accuracy, F1, precision, recall, or AUC) depending on the dataset characteristics and clinical context. Beyond this, ablation experiments will be conducted to systematically evaluate the contribution of each component in the FedPA pipeline, enabling us to identify the most effective configurations for stability and generalization. Ultimately, these enhancements aim to finalize FedPA into a robust, performance-aware federated learning algorithm that is well-suited for real-world commercial deployment in healthcare AI applications.

Author Contributions

Atif Mahmood and Tashin Khan Sadique are the main contributors to the conceptualization, methodology, and manuscript preparation. The remaining authors contributed to data analysis, validation, and review of the manuscript.

Data Availability Statement

The study utilized publicly available datasets. The dataset and corresponding code will be made available upon reasonable request.

References

[1] F. Fatoni, T. B. Kurniawan, D. A. Dewi, M. Z. Zakaria, and A. M. M. Muhayeddin, "Fake vs real image detection using deep learning algorithm," *Journal of Applied Data Sciences*, vol. 6, no. 1, pp. 366–376, 2025.

[2] D. S. Ting, Y. Liu, P. Burlina, X. Xu, N. M. Bressler, and T. Y. Wong, "Ai for medical imaging goes deep," *Nature medicine*, vol. 24, no. 5, pp. 539–540, 2018.

[3] S. Iqbal, A. N. Qureshi, J. Li, and T. Mahmood, "On the analyses of medical images using traditional machine learning techniques and convolutional neural networks," *Archives of Computational Methods in Engineering*, vol. 30, no. 5, pp. 3173–3233, 2023.

[4] A. S. Panayides, A. Amini, N. D. Filipovic, A. Sharma, S. A. Tsaftaris, A. Young, D. Foran, N. Do, S. Golemati, T. Kurc *et al.*, "Ai in medical imaging informatics: current challenges and future directions," *IEEE journal of biomedical and health informatics*, vol. 24, no. 7, pp. 1837–1857, 2020.

[5] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[6] C. A. Gao, N. S. Markov, T. Stoeger, A. Pawlowski, M. Kang, P. Nannapaneni, R. A. Grant, C. Pickens, J. M. Walter, J. M. Kruser *et al.*, "Machine learning links unresolving secondary pneumonia to mortality in patients with severe pneumonia, including covid-19," *The Journal of Clinical Investigation*, vol. 133, no. 12, 2023.

[7] P. Khan, M. F. Kader, S. R. Islam, A. B. Rahman, M. S. Kamal, M. U. Toha, and K.-S. Kwak, "Machine learning and deep learning approaches for brain disease diagnosis: principles and recent advances," *Ieee Access*, vol. 9, pp. 37 622–37 655, 2021.

[8] A. Kilic, "Artificial intelligence and machine learning in cardiovascular health care," *The Annals of thoracic surgery*, vol. 109, no. 5, pp. 1323–1329, 2020.

[9] A. Barragán-Montero, U. Javaid, G. Valdés, D. Nguyen, P. Desbordes, B. Macq, S. Willems, L. Vandewinckele, M. Holmström, F. Löfman *et al.*, "Artificial intelligence and machine learning for medical imaging: A technology review," *Physica Medica*, vol. 83, pp. 242–256, 2021.

[10] W. Salah Eldin and A. Kaboudan, "Ai-driven medical imaging platform: Advancements in image analysis and healthcare diagnosis," *Journal of the ACS Advances in Computer Science*, vol. 14, no. 1, 2023.

[11] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[12] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, "Federated learning and differential privacy for medical image analysis," *Scientific reports*, vol. 12, no. 1, p. 1953, 2022.

[13] M. N. Hossen, K. Ahmed, F. M. Bui, and L. Chen, "Fedrsmax: An effective aggregation technique for federated learning with medical images," *2023 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pp. 229–234, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:264521437

[14] L. Kwak and H. Bai, "The role of federated learning models in medical imaging," *Radiology: Artificial Intelligence*, vol. 5, no. 3, p. e230136, 2023.

[15] T. Li, M. Sanjabi, A. Beirami, and V. Smith, "Fair resource allocation in federated learning," *arXiv preprint arXiv:1905.10497*, 2019.

[16] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[17] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," *Advances in neural information processing systems*, vol. 30, 2017.

[18] H. Zhao, "Non-iid quantum federated learning with one-shot communication complexity," *Quantum Machine Intelligence*, vol. 5, no. 1, p. 3, 2023.

[19] S. Amgain, P. Shrestha, S. Bano, I. del Valle Torres, M. Cunniffe, V. Hernandez, P. Beales, and B. Bhattarai, "Investigation of federated learning algorithms for retinal optical coherence tomography image classification with statistical heterogeneity," *ArXiv*, vol. abs/2402.10035, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267681948

[20] H. Guan, P.-T. Yap, A. Bozoki, and M. Liu, "Federated learning for medical image analysis: A survey," *Pattern Recognition*, p. 110424, 2024.

[21] L. Peng, G. Luo, A. Walker, Z. Zaiman, E. K. Jones, H. Gupta, K. Kersten, J. L. Burns, C. A. Harle, T. Magoc *et al.*, "Evaluation of federated learning variations for covid-19 diagnosis using chest radiographs from 42 us and european hospitals," *Journal of the American Medical Informatics Association*, vol. 30, no. 1, pp. 54–63, 2023.

[22] L. Zhou, M. Wang, and N. Zhou, "Distributed federated learning-based deep learning model for privacy mri brain tumor detection," *arXiv preprint arXiv:2404.10026*, 2024.

[23] A. Makkar and K. Santosh, "Securefed: federated learning empowered medical imaging technique to analyze lung abnormalities in chest x-rays," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 8, pp. 2659–2670, 2023.

[24] Y. Xia, D. Yang, W. Li, A. Myronenko, D. Xu, H. Obinata, H. Mori, P. An, S. A. Harmon, E. B. Turkbey, B. I. Turkbey, B. J. Wood, F. Patella, E. Stellato, G. Carrafiello, A. M. Ierardi, A. L. Yuille, and H. R. Roth, "Auto-fedavg: Learnable federated averaging for multi-institutional medical image segmentation," *ArXiv*, vol. abs/2104.10195, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233324290

[25] M. Delehouzée, X. Lessage, T. Reginster, and S. Mahmoudi, "Performance analysis of aggregation algorithms in cross-silo federated learning for non-iid data," *2024 4th International Conference on Embedded & Distributed Systems (EDiS)*, pp. 74–79, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:274708806

[26] M. Subramanian, V. Rajasekar, S. VE, K. Shanmugavadivel, and P. Nandhini, "Effectiveness of decentralized federated learning algorithms in healthcare: a case study on cancer classification," *Electronics*, vol. 11, no. 24, p. 4117, 2022.

[27] C. Mathew and P. Asha, "Fedprox: Fedsplit algorithm based federated learning for statistical and system heterogeneity in medical data communication," *Journal of Internet Services and Information Security*, vol. 14, no. 3, pp. 353–370, 2024.

[28] M. Muntaqim and T. A. Smrity, "Federated learning framework for brain tumor detection using mri images in non-iid data distributions," *Journal of Imaging Informatics in Medicine*, pp. 1–19, 2025.

[29] Z. Zhou, G. Luo, M. Chen, Z. Weng, and Y. Zhu, "Federated learning for medical image classification: A comprehensive benchmark," *arXiv preprint arXiv:2504.05238*, 2025.

[30] N. Koutsoubis, A. Waqas, Y. Yilmaz, R. P. Ramachandran, M. Schabath, and G. Rasool, "Future-proofing medical imaging with privacy-preserving federated learning and uncertainty quantification: A review," *arXiv preprint arXiv:2409.16340*, 2024.

[31] M. Jiang, Z. Wang, and Q. Dou, "Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 1087–1095.

[32] H. Yuan and T. Ma, "Federated accelerated stochastic gradient descent," *ArXiv*, vol. abs/2006.08950, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:219708203

[33] R. Liu, Y. Cao, M. Yoshikawa, and H. Chen, "Fedsel: Federated sgd under local differential privacy with top-k dimension selection," in *Database Systems for Advanced Applications: 25th International Conference, DASFAA 2020, Jeju, South Korea, September 24–27, 2020, Proceedings, Part I 25*. Springer, 2020, pp. 485–501.

[34] F. Haddadpour and M. Mahdavi, "On the convergence of local descent methods in federated learning," *ArXiv*, vol. abs/1910.14425, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:207757958

Privacy-Preserving Education Data Sharing Scheme Based on Consortium Blockchain

Jiaqi Guo[1], Zhuoran Wang[2], Ningning Liu[3]*
School of Cyberspace Security, Xi'an University of Posts and Telecommunications, Xi'an 710121, China[1,2]
Information Technology Center, Hebei Open University, Shijiazhuang 050080, China[3]

***Abstract*—With the growing emphasis on lifelong education and the rapid expansion of open education platforms, the secure and efficient management and sharing of lifelong learning data have become critical challenges. To address these issues, this paper proposes a Privacy-Preserving Educational Data Sharing (PPEDS) scheme based on blockchain technology. The PPEDS scheme employs attribute-based encryption with hidden attributes to achieve privacy-preserving and fine-grained access control. In addition, it incorporates multi-keyword searchable encryption to enable efficient encrypted data retrieval and combines private and consortium blockchains to ensure data authenticity and integrity across multiple educational institutions. The security analysis demonstrates that the scheme resists potential attacks and ensures confidentiality, access control, and search privacy under a semi-trusted model. Furthermore, performance evaluations conducted on real-world educational datasets show that the proposed scheme achieves efficient encryption, search, and decryption operations, with low computational overhead even in large-scale deployments. Overall, the PPEDS scheme provides a secure, scalable, and practical solution for privacy-preserving data sharing in lifelong education systems.**

Keywords*—Consortium blockchain; access control; secure search; life-long education; data sharing

I. Introduction

Lifelong education integrates various forms of learning that individuals receive at different stages of life and is becoming increasingly important in the rapidly developing knowledge-based society [1]. With the advancement of information technology, open education platforms have become a powerful tool for promoting lifelong learning, providing flexible access to educational resources from different institutions and regions [2], [3]. However, traditional approaches to educational data management are hampered by problems such as isolated data silos, limited transparency, and fragmented information systems [4]. These limitations pose significant obstacles to the recognition and portability of learning achievements, thereby affecting the overall effectiveness of the lifelong learning ecosystems. With the rapid development of lifelong education and open education platforms, the management and sharing of educational data have become increasingly complex and critical. The flow of students between different educational platforms and educational institutions, especially the needs of cross-school elective courses, credit certification, transcript verification, etc., has made cross-platform sharing of educational data an urgent need. Therefore, secure data sharing and privacy protection are essential to promote the healthy development of information technology in lifelong education.

*Corresponding author.

Blockchain technology has the characteristics of decentralization, immutability and transparency [5] and has been widely regarded as a transformative solution for secure data management and sharing. Blockchain can achieve secure storage of data, access control and privacy protection, while improving the transparency and credibility of data sharing. Recent studies have explored various applications in fields such as healthcare and finance, aiming to address challenges in data security, privacy, and cross-institutional data interoperability [6], [7].

Although extensive research has explored secure data sharing based on blockchain in domains such as healthcare, transportation, and the Internet of Things (IoT), relatively few studies have focused specifically on educational data sharing. Unlike other application areas, educational data presents unique challenges. It often involves multiple stakeholders—including students, teachers, educational institutions, and government bodies—with complex relationships and frequent data updates. Additionally, educational data is heterogeneous in format, privacy-sensitive, and typically distributed across different platforms, making secure, traceable, and fine-grained access control particularly difficult. Moreover, practical requirements such as cross-institutional credit recognition, transcript verification, and course-based access control necessitate flexible and scalable sharing mechanisms that are not well addressed by existing solutions. In recent years, some progress has been made toward blockchain-enabled educational systems. Huang et al. [8] developed a blockchain-based education data management system that integrates private and consortium blockchains to enhance data security and traceability. Marouan et al. [9] proposed a blockchain-based framework for managing educational credentials and supporting lifelong learning, emphasizing decentralized trust and certificate verification. However, these works mainly focus on identity and credential management, lacking technical details on fine-grained access control, privacy-preserving queries, and efficient data retrieval. Therefore, there is an urgent need to develop a comprehensive and secure data-sharing scheme tailored for educational environments.

Overall, existing research provides a variety of secure and private data sharing methods, but there is still a lack of specific solutions for educational data sharing, and these solutions have certain limitations, such as the lack of fine-grained access control, efficient user decryption, and practical multi-keyword search. In response to these limitations, we propose a privacy-preserving education data sharing scheme based on consortium blockchain, named PPDES, providing a comprehensive solution for secure and efficient lifelong education data sharing.

The main contributions and novel features of this paper are as follows:

- By integrating private blockchain and consortium blockchain, the PPEDS scheme addresses the limitations of traditional data management, ensuring data legitimacy, traceability, and enabling secure data sharing across educational platforms.
- The PPEDS scheme achieves fine-grained access control through multi-authority CP-ABE, allowing each educational platform to manage user attributes independently, while hidden attributes and user anonymity enhance access security.
- Based on public key searchable encryption, the PPEDS scheme supports efficient multi-keyword search. After searching, the consortium blockchain can perform pre-decryption for users who meet access policies, significantly reducing user computation.

The rest of this paper is organized as follows: Section II discusses the system model, threat model, and design goals. Section III introduces the problem formulation. Section IV details the preliminaries. The PEDD system is given in Section V. Sections VI and VII discuss security and performance analyses, respectively. Finally, Section VIII concludes the paper.

II. Related Work

In recent years, blockchain technology has been widely applied in the field of privacy-preserving data sharing across various domains, including healthcare, IoT, and transportation.

Zheng et al. [10] proposed a scalable privacy-preserving data sharing scheme based on blockchain, which utilizes a (p, t) threshold Paillier cryptosystem to enforce user access control. Li et al. [11] introduced a blockchain-based IoT data sharing scheme with privacy protection and a reward mechanism, where signature exchange is used to implement access control. Li et al. [12] designed a privacy-preserving method for IoT medical systems based on blockchain and lightweight secret sharing, combining decentralization and data confidentiality to facilitate secure sharing among institutions. In another work, Li et al. [13] proposed a privacy-preserving flight operation data sharing scheme using zero-knowledge proofs and proxy re-encryption to realize secure access control.

Although these schemes adopt lightweight access control mechanisms, they generally lack fine-grained expressiveness and exhibit limited efficiency when applied to educational environments involving a large number of users and roles.

To address fine-grained access control, Liang et al. [14] presented a personal data protection framework based on consortium blockchain and attribute-based encryption (ABE), enabling secure data sharing among multiple platforms. However, the issue of searchable encryption was not considered in this work, which is crucial for efficient data access in large-scale systems.

In terms of searchable encryption, Jiang et al. [15] proposed an efficient and privacy-preserving data sharing scheme in intelligent transportation systems, employing Bloom filters to support multi-keyword retrieval. Du et al. [16] designed a blockchain-based searchable encryption scheme using both public and private blockchains, where encrypted indexes reside on the private chain and documents are outsourced to the public chain. Their scheme also incorporates access control to improve query efficiency. Niu et al. [17] presented an attribute-searchable scheme based on CP-ABE and blockchain, supporting attribute hiding and multi-keyword search. Liu et al. [18] developed a privacy-preserving medical data sharing framework based on blockchain and ABE, which supports both fine-grained access control and secure keyword search among medical institutions.

However, most of the above solutions have the problem of low efficiency of user decryption or lack of the ability of attribute authorities to independently manage attributes.

While the above works span various application fields such as medical care, IoT, and transportation, there is relatively little research focused on blockchain-based educational data sharing. To fill this gap, Luo et al. [19] propose a blockchain-based education data management system that leverages the strengths of private and consortium blockchains to enhance system security. However, the scheme lacks basic access control and keyword retrieval. Huang et al. [8] developed a blockchain-based educational data management system that leverages both private and consortium blockchains to enhance security. In [9], Marouan et al. present a blockchain-based framework for educational credentials and lifelong learning. They explore the use of a decentralized ledger to issue, store, and verify educational credentials such as course completions, skills certificates, and learning records in a tamper-resistant manner. Their work emphasizes the trustworthiness and interoperability of credentials across institutions, highlighting how blockchain can mitigate fraud and simplify cross-platform verification. However, their schemes still lack mechanisms for fine-grained access control and efficient encrypted search. These limitations highlight the necessity of a specialized security framework tailored to educational data sharing scenarios. Our PPEDS scheme distinguishes itself by addressing these gaps through multi-authority CP-ABE, hidden attributes, efficient searchable encryption, and dual-chain integration.

III. Problem Formulation

A. System Model

The PPEDS scheme involves a large-scale open education platform composed of multiple provincial and municipal education platforms. Each educational platform has its own cloud server and local client, which is typically operated by teachers of the education platform. The registered students also have certain operational permissions, while other registered users only have the right to search and access some public courses. The transcripts and certificates of students, or some paid courses of teachers are encrypted and stored on the platform server. Each educational platform builds its own private blockchain to store the hashes of encrypted content and the keyword indexes. The private blockchains of all participating educational platforms together constitute a consortium blockchain, which is managed by all platforms for cross-institutional collaboration and data sharing. Each platform uploads part of its data, such as the identifier of the private blockchain and keyword indexes to the consortium blockchain.

The system mainly includes five entities: teachers, students, platform servers, private blockchains, and the consortium blockchain. The system model is illustrated in Fig. 1.

Teachers: The users of each education platform are simply divided into two roles: teachers and students. Teachers are typically the data managers of the platform, responsible for evaluating student assignments, exams, etc., and uploading encrypted transcripts and certificates to the platform server, along with generating corresponding keywords for retrieval. Teachers also need to upload their courses. Public courses do not need to be encrypted, and teachers generate keyword indexes and upload them directly. However, some courses are restricted to users with specific attributes. For example, a course of computer major may require students to satisfy {*"Major": "computer science" AND "Passed course": "C language"*} to access it. Teachers also need to generate hash values and signatures for encrypted transcripts or courses and upload them to the private blockchain to generate and broadcast this new transaction. Other nodes on the private blockchain are responsible for verifying the transaction, and if the verification passes, a new block is generated on the private blockchain.

Students: When students search for their grades or courses, they generate the corresponding search trapdoor using the keys assigned to them during registration and the related keywords. To achieve cross-platform interoperability, the search trapdoor is uploaded to the consortium blockchain, where nodes on the consortium execute the search algorithm. Students can only search and access their own transcripts and certificates, and apart from public courses, they can only access courses that are part of their study plan. Grades between different educational platforms can be mutually recognized. To avoid performing complex decryption operations locally, students can generate a temporary key during searches, allowing the platform server to perform pre-decryption operations for the authorized student. In practice, teachers or other registered users can also perform searches and access data, but they should all follow the access permissions of the data owner.

Platform Server (PS): Platform servers are typically large storage and computing servers in the cloud, which are responsible for storing all courses and related content. After storing encrypted transcripts or courses, the PS extracts the block identifiers from the private blockchain, along with the keyword indexes, and generates a signature to construct a new transaction on the consortium blockchain. Other nodes on the consortium blockchain are responsible for verifying the transaction, and if the verification passes, a new block is generated on the consortium blockchain. When the searched content is encrypted data, PS is responsible for pre-decrypting it for the user before returning the content.

Private Blockchain (PB): Teachers upload the hash values of encrypted transcripts or courses, keyword indexes, and their signatures to the PB, creating a new transaction. Nodes on the PB verify the signature to ensure it was legally generated by the teacher, and then generate a new block on the PB. The data structure of PB is shown in Fig. 2.

Consortium Blockchain (CB): The nodes on the CB are responsible for executing the search algorithm. If the search is successful, the nodes on CB extract the security index from the block to obtain the block identifier of the PB. Using the block identifier from the PBn, they retrieve the ciphertext hash and provide it to the corresponding PS. After confirming hash consistency, the PS uses the temporary key provided by the user to perform pre-decryption on the ciphertext and returns the pre-decrypted content to the consortium blockchain nodes, which then return to the search user. The data structure of CB is shown in Fig. 3.

B. Threat Model

Assuming the educational platform servers and the computers of teachers or students are semi-trusted, which means that they are trusted in most cases but may still be at risk of being compromised. Attackers may eavesdrop on the information being transmitted through communication channels, such as security indexes, encrypted course content, and search trapdoors. Additionally, clients and servers are not allowed to collude to infer the real identity of users.

The PB is managed by multiple administrators within the educational platform. This multi-administrator setting prevents single-point control and enhances the security of the private blockchain. In this semi-trusted environment, each administrator is assumed to be independently trusted, and they do not collude to tamper with data or engage in malicious behaviors.

The CB is managed by multiple educational platforms, with each platform participating in its management and maintenance, sharing part of the data, and achieving cross-platform interconnection. We assume that the nodes of these platforms do not collude with each other to ensure the security and integrity of the consortium blockchain data.

C. Design Goals

Based on the semi-trusted assumptions and potential attack risks described in Section 2.3, the security objectives are as follows:

1) Data security and access control: Sensitive data must be encrypted, and data owners should have the freedom to define access control policies, ensuring that only users who meet the specified criteria can access the data.

2) Privacy protection: Attributes in access control policies must not reveal sensitive user information, and ciphertext should be resistant to DDH testing to ensure recipient anonymity.

3) Secure search: Indexes and queries should remain encrypted, and searches should be performed by designated trusted nodes to prevent attackers from obtaining the true keywords during the search process.

4) System availability: Consistency in the private blockchain is achieved through the verification of signatures from internal administrators, while the consortium blockchain's consistency is maintained through platform signature verification, neither of which involves sensitive information.

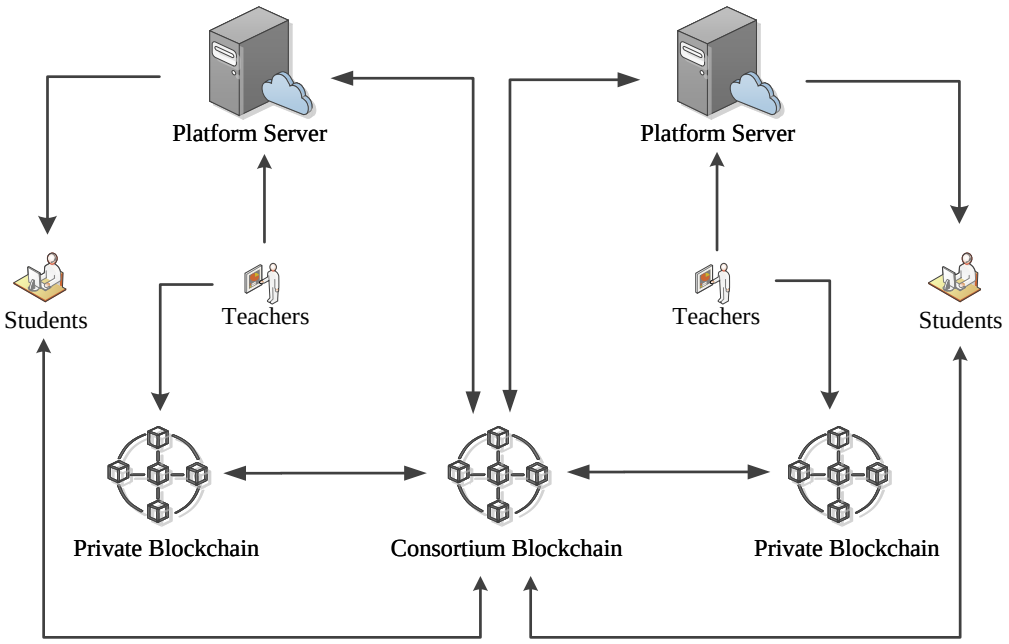


Fig. 1. System model.

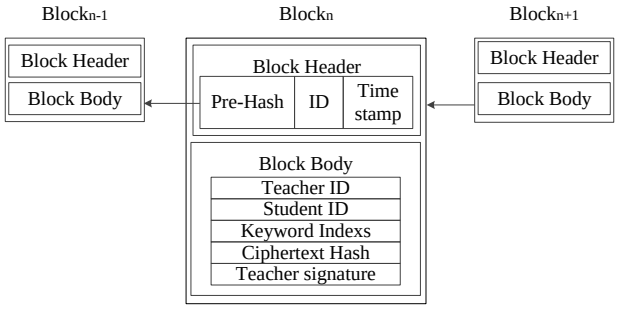


Fig. 2. Data structure of private blockchain.

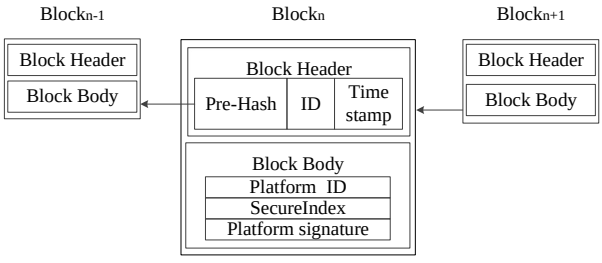


Fig. 3. Data structure of consortium blockchain.

IV. Preliminaries

A. Attribute-Based Encryption

Attribute-Based Encryption (ABE) [20] is a public key encryption scheme that allows users to encrypt and decrypt data based on a set of attributes. The main advantage of ABE is that it enables fine-grained access control and allows flexible data access management. There are two main types of ABE: Key-Policy ABE (KP-ABE) [20] [21] and Ciphertext-Policy ABE (CP-ABE)[22].

- KP-ABE: In KP-ABE, the access policy is embedded in the secret key, while the ciphertext is labeled with attributes.
- CP-ABE: In CP-ABE, the access policy is embedded in the ciphertext, while the secret key is associated with attributes. This allows the data owner to define who can access the encrypted data, making CP-ABE well-suited for fine-grained access control.

Formally, let G be a cyclic group of prime order p, and g be its generator. During the system setup, public parameters and a master key are generated. For each attribute i, a secret key component is generated using a random value. The decryption key is derived based on the attributes they possess. The encryption function takes a message M, an access policy A, and outputs a ciphertext C. The decryption function allows a user to recover the message M if their attributes satisfy the policy A.

B. Searchable Encryption

Searchable Encryption (SE) [23] is a cryptographic Primitive that allows encrypted data to still support keyword-based search. SE enables the platform to store data in encrypted form while allowing users to search over it without revealing sensitive information. The main components of SE include:

1) Index encryption: The data owner encrypts the index keyword w as I_w. The encrypted index I_w is then uploaded to the server.

2) Trapdoor generation: To search for a keyword w', the user generates a trapdoor $T_{w'}$.

3) Search: The server uses the trapdoor T_w to search through the encrypted index $\{I_w\}$. When $w = w'$, the server returns the matching content to the user without learning the specific information of w or w'.

C. Blockchain

Blockchain [24] is a distributed ledger technology that provides data integrity, security, transparency, and immutability through decentralized consensus mechanisms. It allows records to be securely stored in linked blocks, making it nearly impossible to alter data once added.

Blockchain can be classified into three main types: public blockchain, private blockchain, and consortium blockchain. Public blockchains are open to everyone, allowing anyone to participate in the consensus process. Private blockchains are restricted to specific participants, typically managed by a single organization, and provide more control over the data. Consortium blockchains are semi-decentralized, managed by a group of institutions, and are ideal for collaborative projects.

V. The Privacy-Preserving Education Data Sharing (PPEDS) Scheme

The PPEDS scheme specifically includes setup, key generation, encryption, blockchain generation, trapdoor generation, search, and decryption. The specific description of each algorithm is as follows.

A. Setup

The setup of the system is to generate some necessary public parameters. First, it selects a security parameter κ. Then, a multiplicative cyclic group G of prime number p order is selected, where g is its generator. Each user entering the system is assigned an identity ID, which can be composed of any string. The ID can be mapped to G using the function H_1. Each user role has its attributes, such as $\{\text{“}major\text{”} : \text{“}mathematics\text{”}\}$, and they are all expressed in the form of $\{attributename : attributevalue\}$. The function H_1 can also map the attribute value to G. Define a function H_2 to map the attribute back to the authority that manages the attribute. In general, the public parameters include $\{p, G, g, H_1, H_2\}$.

Next, each authority generates a pair of public and private keys. Assume that the authority picks four random numbers $\alpha, \beta, \gamma, \sigma \in Z_p$ as its private key. Then, it calculates $e(g,g)^{\alpha}, g^{\beta}, g^{\gamma}, g^{\sigma}$ as its public key and publishes them.

B. KeyGen

Each authority (i.e., each educational platform) distributes corresponding attribute keys to registered users with different roles (teachers, students, administrators, etc.) in the system.

Suppose an authority needs to generate an attribute key for a student named Alice, whose attribute is $v_\chi \in V_a$, where V_a is Alice's attribute set. Each authority responsible for V_a needs to generate a corresponding attribute key for the user. For example, if Alice is a teacher at platform PS_1 but is also a student in a specific course at platform PS_2, both PS_1 and PS_2 need to generate attribute private keys for Alice. Where ID_a is Alice's identifier, which can be composed of any string. We will use ID_a below to represent a student user like Alice.

$$\begin{aligned} & sk_{ID_a, v_\chi} \\ =&\{sk_1 = g^{\frac{\alpha}{\gamma}} H_1(ID_a)^{\frac{\beta}{\gamma}} H_1(v_\chi)^{\frac{\sigma}{\gamma}}, \\ & sk_2 = H_1(v_\chi)^{\frac{\sigma}{s}}, sk_3 = g^{\sigma}\} \end{aligned} \tag{1}$$

Obviously, all the keys related to $v_\chi \in V_a$ composes the attribute key $sk_a = \{sk_{ID_a,v}\}$. Each successfully registered student will receive a search key s distributed by the system to generate a search trapdoor.

C. Enc

The teacher extracts keywords (such as student name, ID, and subject) from course grades and encrypts them, and also extracts keywords from some non-public courseware and other learning content and encrypts them before uploading them to the platform server. Attribute-based encryption ensures that only students who meet the access policy specified by the teacher can decrypt the content.

1) FileEnc: Suppose the teacher needs to upload Alice's mathematics transcript T_a and only allow Alice and other teachers can access it. For the sake of generalization, assume that this specific access policy is represented by (A, f, Val_p). Here $A \in Z_p^{c \times d}$ is an access matrix, and f is a function that maps rows in A to attribute names. Let $Val_p = v_{f(1)}, \cdots, v_{f(c)}$ represent the corresponding attribute value. Among them, Val_p contains specific attribute values, such as the student's name or ID, which needs to be kept private, and the other parts of the access policy need to be uploaded with the ciphertext.

A function η is also defined to map the rows of A to the authorities, that is, $\eta(\cdot) = H_2(f(\cdot))$. Then, select some random numbers in Z_p and generate two vectors $\vec{a} = (z, a_2, \cdots, a_d)^{\top}$ and $\vec{b} = (0, b_2, \cdots, b_d)^{\top}$. According to LSSS [20], $a_i = A_i \vec{a}$ and $b_i = A_i \vec{b}$, where A_i represents the ith row of A.

For each $i \in [1, c]$ in A, randomly select $r_i, r_i' \in \mathbb{Z}_p$. The encrypted R is as follows:

$$\begin{aligned} & CT_0 = T_a e(g,g)^z, \\ \forall i \in [1,c], \quad & \{CT_{1,i} = g^{a_i} g^{\alpha_{\eta(i)} r_i}, CT_{2,i} = (g^{\gamma})^{-r_i}, \\ & CT_{3,i} = (g^{\delta})^{r_i - r_i'}, CT_{4,i} = g^{\beta_{\eta(i)} r_i} g^{b_i}, \\ & CT_{5,i} = H_1(v_{f(i)})^{r_i'}\} \end{aligned} \tag{2}$$

The ciphertext $CT = \{(A, f), CT_0, \{CT_{1,i}, CT_{2,i}, CT_{3,i}, CT_{4,i}, CT_{5,i}\}_{i \in [1,c]}\}$.

2) IndexEnc: The teacher selects several keywords such as $W = \{Alice, ID_a, Mathematics\}$, and choose a random number $\tau \in Z_p$ and encrypts them as:

$$\begin{aligned} I_1 &= e(g,g)^{\tau} \\ I_2 &= g^{s\tau} \\ I_w &= \{g^{\tau H(w)}\}_{w \in W} \end{aligned} \tag{3}$$

The encrypted index is $I = I_1, I_2, I_w$. Then the ciphertext CT is sent to PS, and the hash value of CT, the encrypted index I, and the signature of the teacher are all uploaded to the PB.

D. BlockGen

Teachers on each education platform upload data to their private chain, and each education platform integrates the data from the private chain into the alliance chain.

1) PriBlockGen: The teacher constructs a new transaction using the keyword index, teacher ID, and teacher signature, then broadcasts the transaction to the PB of the education platform. Upon receiving the new transaction, the verifiers on the PB verify the signature in the transaction. If it is valid and confirms that the content was legally created by an authorized teacher, a verification confirmation message is broadcast. When more than 2/3 of the verifiers confirm, the PB accepts the new transaction and generates a new block. Otherwise, the new block is rejected.

2) ConBlockGen: The platform server extracts the ID and keyword index from each new block of its PB to generate a security index. The PS constructs a new transaction using the security index, platform ID, and platform signature, then broadcasts the transaction to the CB. Upon receiving the new transaction, the verifiers on the CB verify the signature of PS. When more than 2/3 of the verifiers confirm, a new block is then generated, ensuring the legitimacy of the content across platforms.

E. TrapGen

When student ID_a wants to search for something of interest, such as her mathematics course transcript, a trapdoor T_w is generated, where w' is the keyword.

$$T_w = g^{\frac{1}{H(w')+s}} \tag{4}$$

If there is more than one keyword of interest, the trapdoor is generated as:

$$T_w = g^{\frac{1}{\sum_{w'=1}^{n} H(w')+s}} \tag{5}$$

If ID_a wants to get the pre-decrypted search results, she also needs to select a random number $t \in Z_p$ to generate a temporary key sk'_a and upload it with the trapdoor. Specifically, sk'_a is calculated as follows.

$$\begin{aligned} sk'_a =& \{sk'_{ID_a, v_\chi}\} \\ =& \{\{sk'_1 = sk_1^{\frac{1}{t}}, sk'_2 = sk_2^{\frac{1}{t}}, sk'_3 = sk_3^{\frac{1}{t}}\}, \\ & ID' = H_1(ID_a)^{\frac{1}{t}}, g^{\frac{1}{t}}\} \\ =& \{\{sk'_1 = g^{\frac{\alpha}{t\gamma}} H_1(ID_a)^{\frac{\beta}{t\gamma}} H_1(v)^{\frac{\sigma}{t\gamma}}, \\ & sk'_2 = H_1(v)^{\frac{\sigma}{t\delta}}, sk'_3 = g^{\frac{\sigma}{t}}\}, \\ & ID' = H_1(ID_a)^{\frac{1}{t}}, g^{\frac{1}{t}}\} \end{aligned} \tag{6}$$

F. Search

The node on the CB calculates and verifies whether the equation $e(I_2 \cdot \prod_{i=W} I_w, T_w) = I_1$ holds. If so, the Platform ID is obtained. The corresponding PB is accessed through the ID to obtain the hash value that can be used to generate the ciphertext. The platform server compares the ciphertext CT and returns it to the node on the CB. The specific calculation process is as follows.

$$\begin{aligned} & e(I_2 \cdot \prod_{w \in W} I_w, T_w) \\ =& e(g^{s\tau} \cdot \prod_{w \in W} g^{\tau H(w)}, g^{\frac{1}{\sum_{w' \in W} H(w')+s}}) \\ =& e(g,g)^{\frac{\tau(\sum_{w' \in W} H(w)+s)}{\sum_{w' \in W} H(w)+s}} \end{aligned} \tag{7}$$

If $w = w'$, the above equation is $e(g,g)^{\tau} = I_1$. The ciphertext corresponding to the index where the equation holds is the search result.

G. Dec

When the search is completed, if ID_a sends a temporary key for pre-decryption, the node on the CB pre-decrypts the ciphertext and sends it to ID_a for final decryption; otherwise, the search results are directly sent to ID_a for decryption by herself.

1) Pre-Dec: If the attribute value Val_a of ID_a satisfies the access policy (A, f, Val_p) in CT, then CB uses the temporary key sk'_a to help her pre-decrypt CT into $pre\text{-}CT$. Where $i \in minA$ is the minimum set of attributes that satisfy the access policy, and $x_i \in Z_p$ are some constants that satisfy $\sum_{i \in minA} x_i A_i = (1, 0, \cdots, 0)$.

$$\begin{aligned} & \prod_{i \in minA} (e(g^{\frac{1}{t}}, CT_{1,i}) \cdot e(sk'_1, CT_{2,i}) \cdot e(sk'_2, CT_{3,i}) \cdot \\ & e(sk'_3, CT_{5,i}) \cdot e(ID'_a, C_{4,i}))^{x_i} \\ =& \prod_{i \in minA} e(g^{\frac{1}{t}}, g^{a_i} g^{\alpha_{\eta(i)} r_i}) \cdot e(g^{\frac{\alpha}{t\gamma}} H_1(ID'_a)^{\frac{\beta}{t\gamma}} \\ & H_1(v_\chi)^{\frac{\sigma}{t\gamma}}, (g^{\gamma})^{-r_i}) \cdot e(H_1(v_\chi)^{\frac{\sigma}{t\delta}}, (g^{\delta})^{r_i - r'_i}) \cdot \\ & e(g^{\frac{\sigma}{t}}, H_1(a_{f(i)})^{r'_i}) \cdot e(H_1(ID_a)^{\frac{1}{t}}, g^{\beta_{\eta(i)} r'_i} g^{b_i}))^{x_i} \\ =& \prod_{i \in minA} (e(g,g)^{\frac{a_i}{t}} \cdot e(H_1(ID'_a), g)^{\frac{b_i}{t}})^{x_i} \\ =& e(g,g)^{\sum_{i \in minA} \frac{a_i x_i}{t}} \cdot e(H_1(ID'_a), g)^{\sum_{i \in minA} \frac{b_i x_i}{t}} \\ =& e(g,g)^{\frac{z}{t}} \end{aligned} \tag{8}$$

Then, $pre\text{-}CT = e(g,g)^{\frac{z}{t}}$ is sent to ID_a. Otherwise, the blockchain refuses to pre-decrypt for u. However, if Val_a does not satisfy the access policy contained in CT, CB refuses to pre-decrypt for ID_a.

2) Dec: When ID_a obtains the pre-ciphertext from CB, she just does a simple operation as $\frac{CT_0}{(pre\text{-}CT)^t} = \frac{CT_0}{(e(g,g)^{\frac{z}{t}})^t} = T_a$ to get her transcript.

VI. Security Analysis

Based on the security goals defined in Section III-C, we conduct a comprehensive security analysis of the PPEDS scheme.

A. Data Security and Access Control

The PPEDS scheme ensures data security and fine-grained access control through attribute-based encryption. For example, a teacher defines an access policy such as {*"Major": "computer science" AND "Name": "Alice"*}. With the multi-authority CP-ABE, only users named Alice and majoring in computer science can decrypt the data. The server can provide pre-decryption services for students if the specific access policy is met, ensuring that only authorized users receive the necessary data. Users who do not meet the access policy cannot achieve pre-decryption for them. The specific formal proof can be found in [25].

B. Privacy Protection

Privacy protection is achieved by minimizing the exposure of user attributes and ensuring anonymity. The system ensures that the attributes in the access policies do not leak sensitive user information. For example, the access policy is {*"Major": "computer science" AND "Name": "Alice"*}, where {*"Major": "computer science" AND "Name": "Alice"*} is uploaded to the education platform server along with the ciphertext, and the specific major information and name are not displayed in the ciphertext. Therefore, obtaining the ciphertext will not obtain the specific attribute value, thereby protecting user privacy. In addition, the ciphertext can also resist the Decisional Diffie-Hellman test (DDH-test) [26], making it infeasible for attackers to infer information about the recipient through brute-force methods. This ensures that users maintain anonymity during data access, preserving privacy even in a semi-trusted environment.

C. Secure Search

To maintain data confidentiality during searches, the PPEDS scheme uses encrypted indexes and query trapdoors, which are processed by trusted nodes. The secure search mechanism is built on the Identity-Based Encryption (IBE) scheme from [27]. According to the security analysis in [27], the PPEDS scheme also resist Chosen-Keyword Attacks (CKA). By encrypting both indexes and queries, the system prevents attackers from learning sensitive information during the search process. Only trusted nodes execute the search, ensuring that authorized parties can retrieve results while unauthorized entities remain unable to determine the actual keywords involved, thereby maintaining confidentiality.

D. System Availability

System availability is ensured through the use of blockchain-based consistency mechanisms. In the private blockchain, consistency is maintained by verifying the signatures of internal administrators, while in the consortium blockchain, consistency is ensured through platform-level signature verification. These verification mechanisms prevent unauthorized modifications while allowing the system to continue functioning even if some nodes are compromised. The use of consensus protocols helps guarantee that data and system integrity are preserved without leaking sensitive information, ensuring that the system remains available and reliable.

The PPEDS scheme effectively mitigates the risks brought by the semi-trusted environment and ensures data security, privacy protection, secure search, and high system availability. By adopting attribute encryption-based access control, secure search mechanism, and blockchain-based consistency check, the system provides strong protection against unauthorized data access and ensures user privacy even in the face of potential attacks.

VII. Performance Analysis

First, we compare and analyze the characteristic performance of several data sharing schemes in Table I. The comparison shows that only the PPEDS scheme fully supports multi-keyword search, pre-decryption, hidden attributes and decentralized authorities. Its features enhance its applicability in complex scenarios such as lifelong education platforms, which require flexible and secure data sharing.

Then, we analyze the theoretical time complexity of the major algorithms, and use several time-consuming operations to represent them. E represents the exponential operation in the group G, E_T represents the exponential operation in the group G_T, and P represents the bilinear pairing operation. As shown in Table II, n_a represents the number of attributes, and n_k represents the number of keywords. The time complexity of TrapGen, Search, and Dec is nearly fixed. The time complexity of EileEnc and pre-Dec is proportional to the number of attributes in the access policy, while the time complexity in IndexEnc is linearly related to the number of keywords to be encrypted.

TABLE I. Feature Comparison

schemes	F1	F2	F3	F4	F5	F6
[10]	personal data	×	×	×	×	×
[11]	IoT	×	×	×	×	×
[12]	Medical	×	×	×	×	×
[13]	Civil aviation	×	×	×	×	×
[14]	personal data	×	×	×	×	×
[15]	transportation	×	✓	×	×	×
[16]	personal data	×	✓	×	×	×
[17]	Medical	✓	✓	×	✓	×
[18]	Medical	✓	✓	✓	✓	×
[8]	Education	✓	×	×	×	×
PPEDS	Education	✓	✓	✓	✓	✓

Note: F1: Application Background; F2: Fine-grained access control; F3: Keyword search; F4: Pre-decryption; F5: Hidden attributes; F6: Decentralized authorities.

Compared with two typical scheme [17], [18], which are two schemes proposed in the medical field that supports keyword search and attribute hiding, the PPEDS scheme takes longer than the [18] scheme in the FileEnc stage. This is because the PPEDS scheme not only supports hiding attributes, but also supports distributed attribute authorities to manage attributes, and can realize user anonymity. Complex operations bring a safer and more private user experience. The two schemes embed access control into the search process, that is, only users whose attributes meet the policy can search. This results in the search part being secure but too inefficient in actual use, making it unsuitable for sharing educational

TABLE II. TIME COMPLEXITY COMPARISON

Algorithms	FileEnc	IndexEnc	TrapGen	Search	Dec
[17]	-	$(2n_a+6)E+P$	$(2n_a+n_k+2)E$	$(2n_a+n_k+2)P+E$	-
[18]	$E+E_T$	$(2n_a+n_k+1)E+E_T+n_kP$	$(n_a+4)E$	$(n_a+3)P+(n_a+n_k+2)E_T+E$	P
PPEDS	$7n_aE$+E_T	$(n_k+1)E+E_T$	E	P	E_T

- Indicates that this algorithm is not involved.

data. The high efficiency demonstrated by the PPEDS scheme highlights its practicality and suitability in the educational data sharing model.

We test the actual operational efficiency of various algorithms in the PPEDS scheme using Python on a real dataset. The data set is the encrypted transcripts of students majoring in computer science and technology, civil engineering, finance, law, and pharmacy at Hebei Open University in 2020. 100 students are randomly selected from each major, and 10 courses are selected for each major, totaling 5,000 transcripts. All experimental results were obtained using a personal computer with an Intel Core i5-1340P CPU (4.60GHz) and 16GB of memory.

A. FileEnc

Fig. 4 shows the time cost of FileEnc algorithm that the teacher executed using a multi-authority CP-ABE-based algorithm. The horizontal axis represents the number of attributes used in the access policy, ranging from 2 to 10, while the vertical axis represents the time required for encryption. As can be seen from the figure, the time required for file encryption increases linearly with the number of attributes. Although using more attributes in the policy can enhance the granularity and security of access control, it also leads to increased computational overhead. Therefore, a balance must be struck between security requirements and encryption efficiency to avoid performance bottlenecks, especially in scenarios involving a large number of attributes.

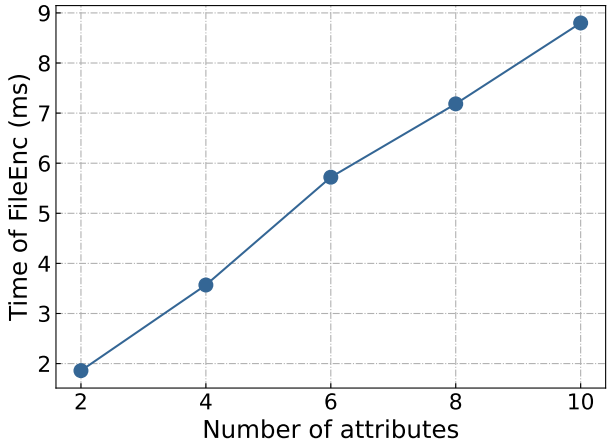


Fig. 4. Time cost for FileEnc.

B. IndexEnc

Fig. 5 presents the time cost of the IndexEnc algorithm. The horizontal axis represents the number of keywords involved in the index, ranging from 2 to 10, while the vertical axis represents the time required for encryption. The figure shows that the encryption time increases linearly as the number of keywords increases. That is, the computational complexity of index encryption is proportional to the number of keywords included in the index. Although adding more keywords can improve data retrieval efficiency and search accuracy, it also leads to increased computational overhead. Therefore, it is crucial to optimize the number of keywords used during indexing to balance search performance and encryption efficiency, especially when dealing with large amounts of encrypted data.

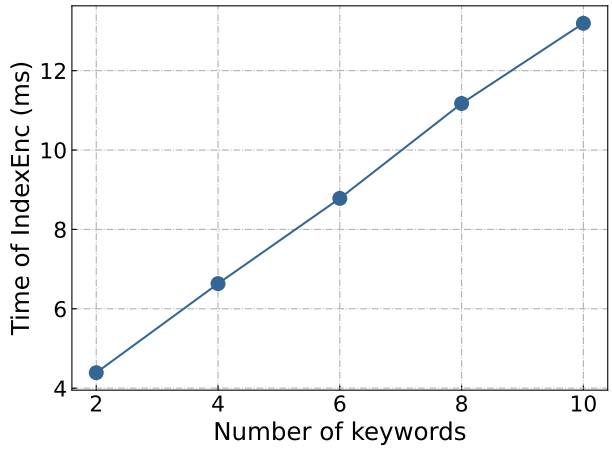


Fig. 5. Time cost for IndexEnc.

C. TrapGen

Fig. 6 illustrates the time cost of the TrapGen algorithm, which represents the time taken to generate a trapdoor for an encrypted search by the searcher. The horizontal axis shows the number of keywords, ranging from 2 to 10, while the vertical axis represents the time required to generate the trapdoor. As can be seen from the figure, the time cost of generating the trapdoor remains relatively constant regardless of the number of keywords involved. This shows that the trapdoor generation process in the PPEDS scheme is efficient and scalable even as the number of keywords increases. The stable performance of the TrapGen operation means that users can search with multiple keywords without significantly increasing the computational overhead. This makes the system well-suited for scenarios where users need to search for multiple attributes or terms simultaneously while maintaining efficient response times. However, too many keywords may result in fewer search results, so the users can adjust them according to their needs.

D. Search

Fig. 7 shows the time cost of executing the search algorithm in the PPEDS scheme. The horizontal axis represents the number of keywords used in the search, ranging from 2 to 10, while

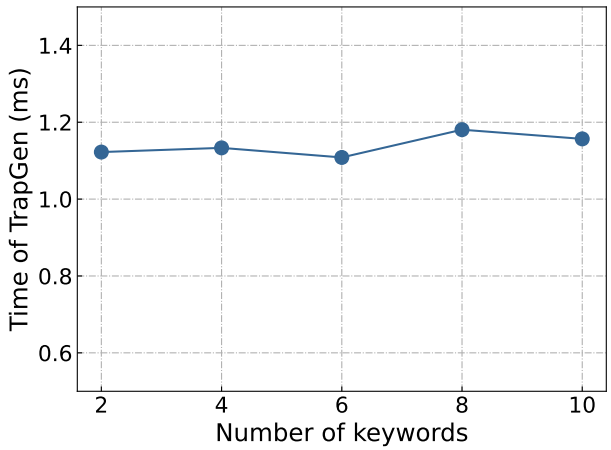


Fig. 6. Time cost for TrapGen.

the vertical axis represents the search time. The figure shows that the time cost of performing the search operation remains relatively constant as the number of keywords increases, and the search process is not significantly affected by the number of search keywords. This consistent performance shows that the proposed search algorithm is both efficient and scalable, even when processing multiple keywords simultaneously. Therefore, the scheme is very suitable for practical applications, and users searching encrypted data using multiple terms will not encounter obvious delays in search efficiency.

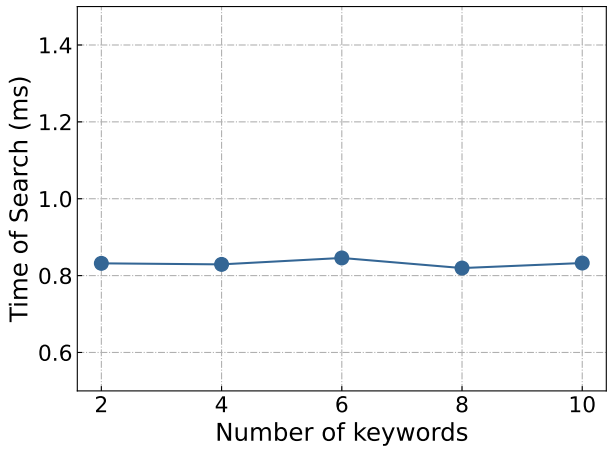


Fig. 7. Time cost for search.

E. Dec

Fig. 8 shows the time cost of the Dec algorithm, comparing the time required for the platform server to perform pre-decryption and the time required for the user to perform the final decryption. The figure includes two graphs: a zoomed-in view to highlight the comparison of the decryption time for the two stages, and a full-scale view that illustrates the overall decryption cost as the number of attributes increases.

The horizontal axis represents the number of attributes involved in decryption, while the vertical axis represents the decryption time. The comparison shows that the pre-Dec performed by the platform server significantly reduces the computational load for the user. In the zoomed-in view, it can be seen that the time cost of the final decryption performed by the user is small and remains almost constant regardless of the number of attributes. In the full-scale graph, the pre-Dec time grows as the number of attributes increases. However, the platform server has powerful computing power and performs the most computationally intensive part of the decryption process, leaving the user with only very small and fixed-time decryption steps. This ensures that the user experiences minimal computational overhead during the decryption process, thereby improving efficiency and maintaining security. By delegating the resource-intensive decryption part to the platform server, the system can strike a balance between user convenience and cryptographic security.

F. Verify by Blockchain

In the PPEDS scheme, the verification process of both private and consortium blockchains relies on the verification of signatures from teachers and educational platforms. We compare three commonly used signature schemes: ECDSA [28], Schnorr [29], and BLS [30] to determine the signature algorithm that best suits different scenarios.

Table III compares the properties of the three schemes, including signature size, speed, and support for aggregation. The ECDSA algorithm has a medium signature size and speed, without support for aggregation, making it suitable for environments where aggregation is not required and moderate signing and verification times are acceptable. Schnorr scheme have a small signature size and support aggregation (e.g., MuSig [31]). BLS scheme has a fixed-size signature and supports native aggregation, which is advantageous in distributed networks like Ethereum 2.0. However, due to high computational cost, BLS is more suitable for scenarios that prioritize aggregation and verification across multiple nodes rather than speed.

Table IV provides the time cost for signing and verification at a security parameter of 256 bits. Overall, for verifying teacher signatures on the private blockchain, Schnorr is preferred due to its small signature size, fast signing and verification times, and support for aggregation. This ensures that the verification process is both efficient and scalable. For the consortium blockchain, where educational platform signatures need to be verified across multiple institutions, BLS signature scheme is recommended. Although BLS has slower signing and verification times, its native support for aggregation makes it suitable for reaching consensus among distributed nodes. This trade-off is acceptable in scenarios where aggregation significantly reduces the number of signatures that need to be processed, thereby improving overall system efficiency.

VIII. Conclusions

This paper proposes the PPEDS scheme, a privacy-preserving data sharing scheme for large-scale, interconnected lifelong education platforms. By integrating private and consortium blockchains, the scheme enables secure, verifiable, and efficient cross-institutional data sharing while preserving user privacy.

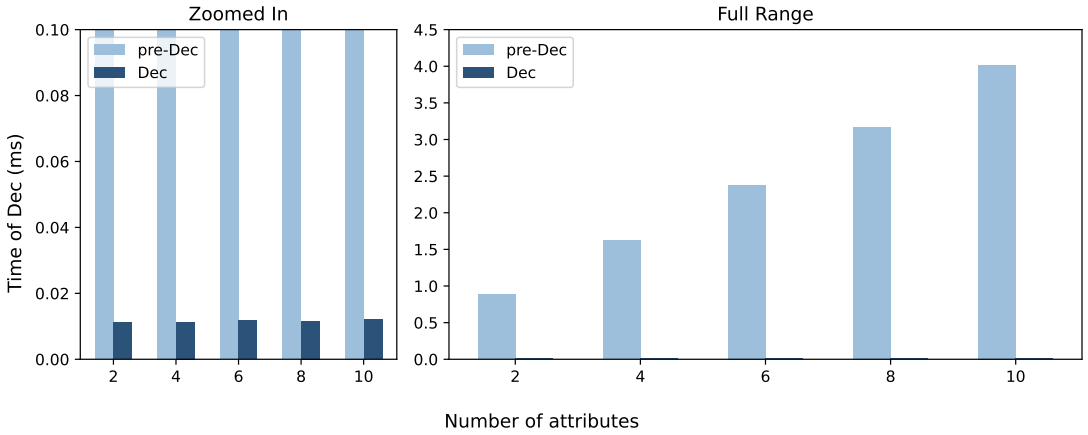


Fig. 8. Time cost for Dec.

TABLE III. FEATURE COMPARISON OF THREE SIGNATURE ALGORITHMS

Schemes	Signature Size	Speed	Aggregation Support	Application Scenarios
ECDSA	Medium	Medium	×	Bitcoin, SSL, TLS
Schnorr	Small	Fast	✓(supports MuSig)	Bitcoin (Taproot), Blockchain
BLS	Fixed-size	Slow	✓(native support)	Ethereum 2.0, Distributed networks

TABLE IV. PERFORMANCE ANALYSIS OF THREE SIGNATURE SCHEMES

Schemes	Sign (ms)	Verify (ms)
ECDSA	0.0903	0.3214
Schnorr	0.0051	0.0078
BLS	4.0647	4.5202

The PPEDS scheme introduces a multi-authority CP-ABE mechanism with hidden attributes to achieve fine-grained access control and user anonymity, allowing each authority to independently manage attribute policies. Additionally, the scheme incorporates public key-based multi-keyword searchable encryption with pre-decryption support via the consortium blockchain, significantly reducing computational overhead for end-users. These innovations collectively enhance the scalability, security, and privacy of educational data sharing.

From a practical perspective, PPEDS addresses real-world challenges in lifelong education systems, such as cross-platform credit recognition, transcript verification, and elective course access. The performance evaluations and comparisons results demonstrate that our pre-decryption mechanism reduces user-side decryption time to only 0.01ms compared to traditional CP-ABE-based systems. The trapdoor generation and search operations maintain sublinear growth with respect to the number of keywords, supporting real-time query capabilities even on resource-constrained devices. In addition, the encryption and index generation overheads remain stable as the number of access attributes increases, showcasing the scalability of the scheme. These metrics highlight the advantage of PPEDS in enabling efficient, scalable, and privacy-preserving data sharing for large-scale educational systems.

However, the current design assumes a semi-honest authority and does not account for collusion among multiple entities. Furthermore, while pre-decryption significantly improves search efficiency, the overhead of storing indexes on the blockchain remains a challenge, potentially impacting scalability over time.

Future work could consider integrating dynamic policy updates and revocation mechanisms to improve flexibility, and optimizing blockchain storage by compressing or pruning index structures to enhance scalability. Furthermore, exploring practical deployments and integration with decentralized identity frameworks will validate the practicality of this solution in real-world educational settings.

ACKNOWLEDGMENT

This work was supported by the Hebei Province Education Science "14th Five Year Plan Project" of 2023 (2303010) and University-level Research Project of Hebei Open University (ZD202203). The authors gratefully acknowledge the anonymous reviewers for their valuable comments.

REFERENCES

[1] J. Field, "Lifelong education," *International Journal of lifelong education*, vol. 20, no. 1-2, pp. 3–15, 2001.

[2] A. Kuznetsov, E. Pyanykh, and M. Rodaikina, "Digital transformation in the context of improving the quality of lifelong education," 2021.

[3] S. F. M. Y. S. Salim, M. F. Mahmood, and A. B. Ahmad, "The importance of information literacy to support lifelong learning in convergence era," *International Journal of Academic Research in Progressive Education and Development*, vol. 7, no. 3, 2018.

[4] S. JANTHAPASS, N. CHANTHAPASSA, and S. KENAPHOOM, "The evolution of lifelong learning: From traditional classrooms to anywhere, anytime education," *Asian Education and Learning Review*, vol. 2, no. 1, pp. 42–54, 2024.

[5] J. Leng, M. Zhou, J. L. Zhao, Y. Huang, and Y. Bian, "Blockchain security: A survey of techniques and research directions," *IEEE Transactions on Services Computing*, vol. 15, no. 4, pp. 2490–2510, 2020.

[6] M. Krichen, M. Ammi, A. Mihoub, and M. Almutiq, "Blockchain for modern applications: A survey," *Sensors*, vol. 22, no. 14, p. 5274, 2022.

[7] N. Deepa, Q.-V. Pham, D. C. Nguyen, S. Bhattacharya, B. Prabadevi, T. R. Gadekallu, P. K. R. Maddikunta, F. Fang, and P. N. Pathirana, "A survey on blockchain for big data: Approaches, opportunities, and future directions," *Future Generation Computer Systems*, vol. 131, pp. 209–226, 2022.

[8] X. Huang, Y. Wang, H. Liang, Y. Ding, Q. Wu, Z. Zhang, and Q. Qu, "Educhain: A blockchain-based privacy-preserving lifelong education platform," in *International Conference on Database Systems for Advanced Applications*. Springer, 2023, pp. 701–706.

[9] A. Marouan, M. Badrani, N. Kannouf, and A. Chetouani, "Empowering education: leveraging blockchain for secure credentials and lifelong learning," in *Blockchain Transformations: Navigating the Decentralized Protocols Era*. Springer, 2024, pp. 1–14.

[10] B.-K. Zheng, L.-H. Zhu, M. Shen, F. Gao, C. Zhang, Y.-D. Li, and J. Yang, "Scalable and privacy-preserving data sharing based on blockchain," *Journal of Computer Science and Technology*, vol. 33, pp. 557–567, 2018.

[11] T. Li, H. Wang, D. He, and J. Yu, "Blockchain-based privacy-preserving and rewarding private data sharing for iot," *IEEE Internet of Things Journal*, vol. 9, no. 16, pp. 15 138–15 149, 2022.

[12] C. Li, M. Dong, X. Xin, J. Li, X.-B. Chen, and K. Ota, "Efficient privacy-preserving in iomt with blockchain and lightweight secret sharing," *IEEE Internet of Things Journal*, 2023.

[13] X. Li, H. Zhao, and W. Deng, "Bfod: Blockchain-based privacy protection and security sharing scheme of flight operation data," *IEEE Internet of Things Journal*, 2023.

[14] W. Liang, Y. Yang, C. Yang, Y. Hu, S. Xie, K.-C. Li, and J. Cao, "Pdpchain: A consortium blockchain-based privacy protection scheme for personal data," *IEEE Transactions on Reliability*, vol. 72, no. 2, pp. 586–598, 2022.

[15] S. Jiang, J. Cao, H. Wu, K. Chen, and X. Liu, "Privacy-preserving and efficient data sharing for blockchain-based intelligent transportation systems," *Information Sciences*, vol. 635, pp. 72–85, 2023.

[16] R. Du, C. Ma, and M. Li, "Privacy-preserving searchable encryption scheme based on public and private blockchains," *Tsinghua Science and Technology*, vol. 28, no. 1, pp. 13–26, 2022.

[17] S. Niu, M. Song, L. Fang, F. Yu, S. Han, and C. Wang, "Keyword search over encrypted cloud data based on blockchain in smart medical applications," *Computer Communications*, vol. 192, pp. 33–47, 2022.

[18] J. Liu, Y. Fan, R. Sun, L. Liu, C. Wu, and S. Mumtaz, "Blockchain-aided privacy-preserving medical data sharing scheme for e-healthcare system," *IEEE Internet of Things Journal*, vol. 10, no. 24, pp. 21 377–21 388, 2023.

[19] L.-j. Luo, H.-b. Lei[1], and Z. Wang, "Educhain: Blockchain-based informative platform for vocational education and training," in *Proceedings of the 2023 4th International Conference on Big Data and Informatization Education (ICBDIE 2023)*, vol. 178. Springer Nature, 2023, p. 454.

[20] A. Sahai and B. Waters, "Fuzzy identity-based encryption," in *Advances in Cryptology–EUROCRYPT 2005: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Aarhus, Denmark, May 22-26, 2005. Proceedings 24*. Springer, 2005, pp. 457–473.

[21] M. Chase, "Multi-authority attribute based encryption," in *Theory of Cryptography: 4th Theory of Cryptography Conference, TCC 2007, Amsterdam, the Netherlands, February 21-24, 2007. Proceedings 4*. Springer, 2007, pp. 515–534.

[22] J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-policy attribute-based encryption," in *2007 IEEE symposium on security and privacy (SP'07)*. IEEE, 2007, pp. 321–334.

[23] D. X. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in *Proceeding 2000 IEEE symposium on security and privacy. S&P 2000*. IEEE, 2000, pp. 44–55.

[24] H. Guo and X. Yu, "A survey on blockchain technology and its security," *Blockchain: research and applications*, vol. 3, no. 2, p. 100067, 2022.

[25] J. Guo, C. Tian, X. Lu, L. Zhao, and Z. Duan, "Multi-keyword ranked search with access control for multiple data owners in the cloud," *Journal of Information Security and Applications*, vol. 82, p. 103742, 2024.

[26] X. Li, D. Gu, Y. Ren, N. Ding, and K. Yuan, "Efficient ciphertext-policy attribute based encryption with hidden policy," in *Internet and Distributed Computing Systems: 5th International Conference, IDCS 2012, Wuyishan, Fujian, China, November 21-23, 2012. Proceedings 5*. Springer, 2012, pp. 146–159.

[27] C. Gentry, "Practical identity-based encryption without random oracles," in *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*. Springer, 2006, pp. 445–464.

[28] D. Johnson, A. Menezes, and S. Vanstone, "The elliptic curve digital signature algorithm (ecdsa)," *International journal of information security*, vol. 1, pp. 36–63, 2001.

[29] C.-P. Schnorr, "Efficient identification and signatures for smart cards," in *Advances in Cryptology—CRYPTO'89 Proceedings 9*. Springer, 1990, pp. 239–252.

[30] D. Boneh, B. Lynn, and H. Shacham, "Short signatures from the weil pairing," in *International conference on the theory and application of cryptology and information security*. Springer, 2001, pp. 514–532.

[31] G. Maxwell, A. Poelstra, Y. Seurin, and P. Wuille, "Simple schnorr multi-signatures with applications to bitcoin," *Designs, Codes and Cryptography*, vol. 87, no. 9, pp. 2139–2164, 2019.

RT-DETR Edge Deployment: Real-Time Detection Transformer for Distracted Driving Detection

Fares Hamad Aljahani
Department of Information Systems-Faculty of Computing and Information Technology,
Northern Border University, Rafha 91911, Saudi Arabia

***Abstract*—Distracted driving is one of the primary contributors to road accidents worldwide, highlighting the urgent need for reliable in-cabin driver monitoring systems. Existing approaches often face trade-offs: CNN-based classifiers achieve high recognition accuracy but lack spatial localization, while lightweight real-time detectors sacrifice contextual reasoning for efficiency. To bridge this gap, we propose a customized fine-tuned transformer-based object detection framework, RT-DETR-L, specifically adapted for distracted driving detection. In contrast to prior applications of RT-DETR, our adaptation integrates distraction-specific data augmentation, loss-balancing strategies, and deployment-oriented optimizations, enabling precise classification and spatial localization of distractions such as texting, drinking, yawning, and eye closure. Trained and validated on a large-scale annotated in-cabin dataset, RT-DETR-L achieves state-of-the-art performance with a mAP50 of 0.995 and mAP50–95 of 0.774. In addition the proposed model demonstrates the deployment feasibility on resource-constrained embedded platforms (ARM-based edge AI devices), where the model sustains real-time performance at 17.5 FPS with minimal latency. These results establish RT-DETR-L as a hybrid solution combining the semantic depth of transformers with the efficiency required for Advanced Driver Assistance Systems (ADAS). By addressing both accuracy and deployability, this study makes concrete contributions toward advancing robust, real-time driver monitoring for enhanced road safety.**

Keywords*—RT-DETR; real-time inference; autonomous vehicles

I. Introduction

Road safety remains a critical challenge worldwide, with millions of accidents occurring annually due to human error and unsafe driving practices. Among the most pressing issues, distracted driving has emerged as one of the leading causes of road crashes, injuries, and fatalities. The increasing prevalence of mobile device usage, in-vehicle infotainment systems, and driver multitasking behaviors has made distraction detection a vital component of modern intelligent transportation systems (ITS). To address this challenge, researchers and practitioners are progressively turning to advanced artificial intelligence (AI), computer vision, and deep learning methods to monitor driver behavior and ensure safer roads.

Recent research has emphasized the multifaceted nature of road safety. Mustapha et al. [1] reviewed advancements in traffic simulation, underscoring the role of predictive modeling in mitigating risks, while Papadimitriou et al. [2] introduced the Road-safety-II framework, highlighting systemic and behavioral barriers to enhanced safety visions. Similarly, Festag et al. [3] stressed the importance of vehicle-to-vehicle (V2V) and roadside sensor communication in achieving proactive safety mechanisms, whereas Alparslan et al. [4] explored how novel engineering materials contribute to safer transportation infrastructures. These works indicate that road safety is a multi-layered problem, where infrastructure, communication, and human behavior converge.

The evolution of connected and autonomous vehicles (CAVs) has further highlighted the necessity of intelligent safety frameworks. Malinverno et al. [5] proposed an edge-based framework to improve the safety of connected cars, demonstrating how real-time analytics at the network edge can enhance responsiveness. Zhao et al. [6] developed TSD-YOLO, a traffic sign detection model ensuring robust recognition in autonomous driving scenarios, while Sabir et al. [7] demonstrated YOLO-based CNN architectures to improve vehicle perception in autonomous platforms. Likewise, Al-Qaness et al. [8] presented an enhanced YOLO-based traffic monitoring system, and Charef et al. [9] applied YOLO to automated traffic violation detection. These studies highlight the growing trend of leveraging YOLO and its variants in traffic-related vision applications.

Nevertheless, while numerous works focus on traffic monitoring and vehicle detection, fewer concentrate explicitly on driver distraction. Several YOLO-based approaches have been designed for robust vehicle perception. For example, Zhu et al. [10] introduced MME-YOLO, a multi-sensor, multi-level enhanced YOLO model for vehicle detection in traffic surveillance, while Liu et al. [11] proposed BGS-YOLO, a YOLOv8-based approach for intelligent target monitoring. These efforts reveal the adaptability of YOLO to diverse road safety applications, yet they do not directly target distracted driving behavior.

In contrast, driver distraction detection has become a specialized focus of recent years. Shen et al. [12] introduced StarDL-YOLO, a lightweight YOLO-based algorithm capable of detecting distracted driving behaviors with reduced computational complexity, suitable for real-time applications. Similarly, Poon et al. [13] explored YOLO-based deep learning networks for distraction detection, demonstrating promising detection rates in compliance engineering contexts. Tanaka et al. [14] compared multiple YOLO-based models for distracted driving detection, highlighting trade-offs between accuracy, latency, and hardware efficiency. Sajid et al. [15] proposed an efficient deep learning framework for distracted driver detection using deep CNNs, while Salakapuri et al. [16] advanced the field further with an integrated deep learning framework combining driver distraction detection and real-time road object recognition within advanced driver-assistance systems (ADAS). Together, these contributions underline the

significance of developing accurate, robust, and computationally efficient distracted driving detection models.

Despite these advancements, existing methods face several limitations. First, many YOLO-based approaches, while effective, rely on convolutional backbones that may not optimally capture global context or multi-scale relationships, especially under complex distraction scenarios involving subtle facial or hand movements. Second, lightweight models often trade off accuracy for efficiency, making them unsuitable for deployment in safety-critical environments where false negatives are unacceptable. Third, comparative studies [14] reveal inconsistencies in performance across datasets, highlighting the lack of generalizability of current architectures. Moreover, while integrated frameworks such as that of Salakapuri et al. [16] combine object recognition with distraction monitoring, they remain constrained by traditional detection paradigms and lack the adaptability of transformer-based architectures.

To overcome these challenges, this paper proposes a novel framework for distracted driving detection built upon the Real-Time Detection Transformer (RT-DETR). Unlike conventional CNN-based models, RT-DETR integrates transformer-based attention mechanisms that capture global dependencies while maintaining real-time efficiency. By leveraging this architecture, our framework aims to balance speed, accuracy, and robustness, enabling reliable distraction detection across diverse driving conditions. Furthermore, the proposed system is designed for seamless integration into ADAS, offering an intelligent and scalable solution to enhance road safety.

The rest of this paper is organized as follows: Section II presents the related work. Section III details the proposed methodology. Section IV reports the experimental results. Section V highlights the detection outcomes. In Section VII, a comparative study is conducted. Section VI presents the results of fine-tuned RT-DETRL deployment on edge devices. Lastly, Section VIII concludes the paper and introduces the future work.

II. Related Work

The literature on road safety, intelligent transportation, and driver distraction spans several domains, including infrastructure development, connected vehicle systems, traffic perception, and driver behavior monitoring. This section reviews and critically discusses prior work, categorized into three major directions: (i) infrastructure and communication for road safety, (ii) vehicle and traffic perception using YOLO-based models, and (iii) distracted driving detection approaches.

A. Infrastructure and Communication for Road Safety

Mustapha et al. [1] highlighted the importance of traffic simulation for predictive road safety, showing how simulation tools can model accident scenarios and evaluate mitigation strategies. However, simulation-based studies often remain abstract and may lack integration with real-world sensing frameworks. Papadimitriou et al. [2] provided the Road-safety-II perspective, which advocates for systemic safety beyond traditional interventions. While visionary, such frameworks face practical barriers, including behavioral resistance and limited implementation scalability. Festag et al. [3] contributed to V2V and roadside communication systems, which enable cooperative awareness among vehicles. Despite their potential, these systems depend heavily on reliable communication infrastructure, which may not be feasible in all regions. Alparslan et al. [4] discussed additive manufacturing of materials for safer roads and vehicles, but their contribution is largely material-science oriented and does not directly address behavioral aspects such as distraction. Together, these works underline the multifactorial nature of road safety, yet none directly confront the issue of driver distraction detection.

B. YOLO-Based Models for Vehicle and Traffic Perception

Several works have leveraged YOLO to enhance perception in traffic environments. Zhao et al. [6] developed TSD-YOLO for robust traffic sign detection, ensuring reliability in autonomous driving. Similarly, Sabir et al. [7] used YOLO-based CNNs for autonomous vehicle safety, while Al-Qaness et al. [8] improved traffic monitoring systems with YOLO. Charef et al. [9] applied YOLO to automated traffic violation detection, and Zhu et al. [10] proposed MME-YOLO, which integrates multi-sensor data for robust vehicle detection. More recently, Liu et al. [11] introduced BGS-YOLO, leveraging YOLOv8 for intelligent road target monitoring. Although these studies showcase the adaptability of YOLO to traffic-related applications, their primary focus is on external traffic entities rather than internal driver states. Thus, while effective in vehicle and infrastructure monitoring, they do not directly address distraction-related risks.

C. Distracted Driving Detection Approaches

Research specifically targeting distracted driving has accelerated in recent years. Shen et al. [12] introduced StarDL-YOLO, a lightweight model balancing detection accuracy and efficiency, but its reduced complexity risks overlooking subtle distractions. Poon et al. [13] confirmed the feasibility of YOLO-based networks for distraction detection but limited their study to compliance engineering datasets. Tanaka et al. [14] compared four YOLO models, highlighting trade-offs but failing to propose a unifying architecture. Sajid et al. [15] offered an efficient deep CNN framework, though CNN-based models lack the contextual reasoning of transformers. Salakapuri et al. [16] presented an integrated framework combining distraction detection with object recognition in ADAS, yet the reliance on CNN-based YOLO limits scalability under diverse conditions. Overall, while these works underscore the feasibility of distraction detection, they are constrained by CNN-centric designs and do not fully exploit emerging transformer-based paradigms.

D. Research Gap and Our Contribution

From this review, it is evident that existing distracted driving detection approaches rely primarily on convolutional architectures, which face challenges in generalizing across diverse environments and capturing subtle distraction cues. Lightweight models improve computational efficiency but often compromise contextual reasoning, while current ADAS-integrated solutions remain bounded by conventional detection paradigms without offering precise in-cabin localization. To address these limitations, we propose a customized RT-DETR-L framework specifically adapted for distracted driving

detection. Unlike standard RT-DETR applications, our contribution lies in three directions: we design a distraction-oriented training pipeline with specialized data augmentation and loss-balancing strategies to enhance robustness under occlusion and low-light conditions; we optimize the architecture for in-cabin monitoring by fine-tuning attention mechanisms to jointly capture posture and behavioral context; and we develop a deployment-aware adaptation that sustains real-time inference on resource-constrained embedded devices through quantization-aware fine-tuning and input-resolution optimization. Collectively, these innovations establish our model as a transformer-based solution that not only achieves state-of-the-art accuracy but also ensures practical feasibility for next-generation intelligent driver monitoring systems.

III. RT-DETRL Methodology for Distracted Driving Detection

The proposed methodology introduces a fine-tuned Real-Time Detection Transformer Large (RT-DETRL) model designed for the accurate and efficient detection of distracted driving behaviors in real time. As illustrated in Fig. 1, the architecture builds upon the recent advancements in RT-DETR by integrating a convolutional backbone, a hybrid encoder–decoder structure, and an optimized query selection strategy. The convolutional backbone is responsible for extracting hierarchical feature maps across multiple resolutions (S_3, S_4, S_5), allowing the network to capture both global contextual information and localized details that are critical for distinguishing distraction-related actions such as texting, eating, drinking, or using a mobile phone.

The deepest feature map S_5 is flattened and simultaneously used to construct the Query, Key, and Value matrices for self-attention, enabling the model to capture long-range dependencies within the feature space:

$$\mathcal{Q} = \mathcal{K} = \mathcal{V} = \text{Flatten}(S_5). \tag{1}$$

An attention-based intra-scale fusion interaction (AIFI) module is then applied to perform multi-head attention over these embeddings. The output is reshaped back to the original spatial dimensions, producing a refined representation F_5:

$$F_5 = \text{Reshape}(\text{AIFI}(\mathcal{Q}, \mathcal{K}, \mathcal{V})). \tag{2}$$

To fully exploit hierarchical information, the refined deep feature map F_5 is fused with intermediate-scale features S_3 and S_4 through a Cross-Scale Convolutional Fusion (CCFF) module (see Fig. 2). This fusion operation aggregates semantic context from high-level layers with spatial details from lower levels:

$$\mathcal{O} = \text{CCFF}(\{S_3, S_4, F_5\}), \tag{3}$$

where $\mathcal{O}$ represents the hybrid output encoding. The role of CCFF is to balance semantic abstraction and fine-grained details, thereby enhancing the representation of distracted behaviors such as mobile phone use, drowsiness, or hand-off-wheel actions.

Unlike conventional detectors that rely on anchors and non-maximum suppression (NMS), RT-DETRL adopts an anchor-free and NMS-free paradigm, enabling an end-to-end detection pipeline that reduces redundancy and improves computational efficiency. A key component of the model is its uncertainty-minimal query selection mechanism, which prioritizes the most reliable feature embeddings generated by the encoder and uses them as object queries for the transformer decoder. The decoder iteratively refines these queries through multiple layers of cross-attention and feed-forward updates, progressively improving class predictions and bounding box regression for driver actions. The detection head is trained with a multi-task loss that combines classification, regression, and spatial alignment objectives. The total loss is formulated as:

$$\mathcal{L}_{total} = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{box}\mathcal{L}_{box} + \lambda_{giou}\mathcal{L}_{giou}, \tag{4}$$

where $\mathcal{L}_{cls}$ represents the cross-entropy classification loss, $\mathcal{L}_{box}$ denotes the ℓ_1 regression loss for bounding box coordinates, and $\mathcal{L}_{giou}$ is the Generalized IoU loss used to improve spatial alignment. Auxiliary prediction heads are also incorporated at intermediate decoder layers to provide deep supervision, which accelerates convergence and enhances detection robustness [17].

Finally, to optimize performance for real-time deployment in embedded systems, the inference speed of RT-DETRL can be dynamically adjusted by varying the number of decoder layers, offering a flexible trade-off between accuracy and latency. Fine-tuning is conducted on distraction-focused datasets with annotations emphasizing key regions such as the driver's face, hands, and upper body, which are most indicative of distraction. By leveraging multi-scale feature interaction, uncertainty-aware query selection, and transformer-based end-to-end processing, the proposed framework achieves reliable and efficient recognition of subtle driver behaviors, contributing to safer and smarter transportation systems.

IV. Experiments and Results

A. RT-DETRL Configuration Summary

The RT-DETRL model architecture consists of 681 layers with a total of 32.8M parameters and 108.0 GFLOPs. Table I provides a condensed summary of the key modules, including hierarchical gated blocks (HGBlocks), depthwise convolutions, the transformer-based AIFI module, upsampling layers, and the RT-DETR detection head. Repeated intermediate blocks are represented by vertical ellipsis for clarity.

TABLE I. Condensed Configuration Summary of RT-DETRL

Index	Module	Arguments	Parameters
0	HGStem	[3, 32, 48]	25,248
1	HGBlock	[48, 48, 128, 3, 6]	155,072
2	DWConv	[128, 128, 3, 2, 1, False]	1,408
3	HGBlock	[128, 96, 512, 3, 6]	839,296
⋮	⋮	⋮	⋮
28	RTDETRDecoder	[13, 256, 256, 256]	7,328,567
Total		Layers: 681	32.8M params, 108 GFLOPs

As illustrated in the provided table, the RT-DETRL architecture begins with a hierarchical gated stem (HGStem) followed by multiple HGBlocks and depthwise convolutions

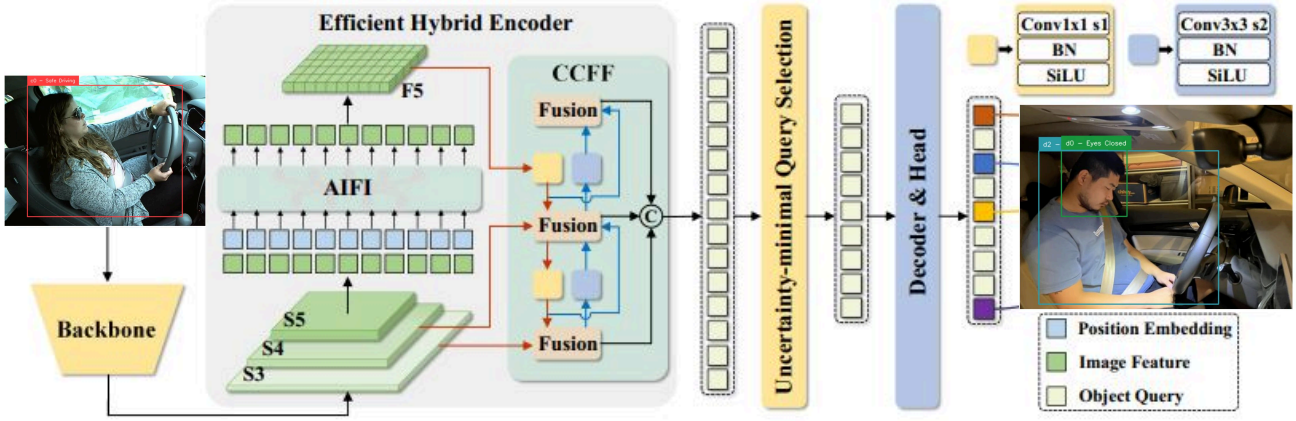


Fig. 1. RT-DETRL architecture for distracted driving detection.

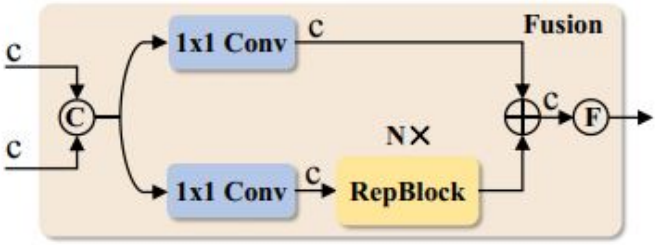


Fig. 2. Cross-Scale Convolutional Fusion (CCFF) unit.

to extract rich multi-scale features. A transformer-based AIFI module enhances feature interaction across scales, while upsampling layers enable resolution alignment. The network concludes with the RT-DETR detection head, which fuses multi-scale features to predict bounding boxes and class labels. Despite its larger size (32.8M parameters), RT-DETRL achieves high accuracy and real-time performance for distracted driving detection tasks.

B. Distracted Driving Dataset

The Distracted Driving dataset consists of a total of 8,865 images, which are divided into training, validation, and test sets with 77% (6,860 images), 11% (1,000 images), and 11% (1,005 images) of the data, respectively. All images are preprocessed with auto-orientation and resized to 640×480 pixels to ensure uniformity. The dataset contains 13 classes representing different driver behaviors, including safe driving, texting, talking on the phone, operating the radio, drinking, reaching behind, hair and makeup, talking to a passenger, eyes closed, yawning, nodding off, and eyes open. This dataset has a wide range of applications: it can be integrated into road safety monitoring systems to detect distracted behaviors in real time, used by companies developing advanced driver assistance systems (ADAS) to enhance driver behavior understanding, employed by insurance companies to evaluate risk and influence policy pricing, leveraged by researchers to study distraction prevalence and inform safety policies, and applied in driver education to promote awareness and safe driving habits [18]. Fig. 3 illustrates the dataset samples.

1) Dataset distribution: The correlogram analysis, illustrated in Fig. 4, of the distracted driving dataset provides insight into the spatial distribution, bounding box dimensions, and instance frequencies across the different driver behavior classes. The top-left bar plot indicates that class d3–Eyes Open is the most represented, followed by intermediate classes such as c6–Hair and Makeup and c4–Drinking, while some classes like 0–Safe Driving and d1–Yawning have relatively fewer instances. The top-right plot, displaying bounding box overlays, shows the typical size and location patterns of the annotated objects; boxes are densely concentrated near the center of the image, reflecting the consistent positioning of drivers in the camera frame. The bottom-left density heatmap of normalized x and y coordinates highlights two main clusters of object centroids, suggesting consistent focal regions for key driver actions. Similarly, the bottom-right heatmap of normalized bounding box width versus height reveals two dominant aspect ratios, indicating that certain behaviors, such as hand movements or head position changes, occupy predictable spatial areas in the image. Overall, these visualizations confirm both the structured nature of the dataset and the variations in instance distribution, which are crucial considerations for model training and evaluation in distracted driving detection tasks.

2) Dataset correlogram: The correlogram presented in Fig. 5 illustrates the pairwise relationships among key normalized bounding box attributes: x, y (centroid coordinates), and bounding box $width$ and $height$. Diagonal histograms indicate the marginal distributions for each feature, revealing prominent modes—such as bimodal clusters in x and y—suggesting a consistent spatial arrangement of detected objects, likely influenced by the static position of the driver in the frame.

Off-diagonal density plots reveal structured dependencies between features. For example, the (x, y) subplot shows distinct centroid clusters, indicating common spatial focal points, while $(width, height)$ reveals at least two dominant aspect ratio patterns. The $(x, width)$ and $(y, height)$ pairings suggest

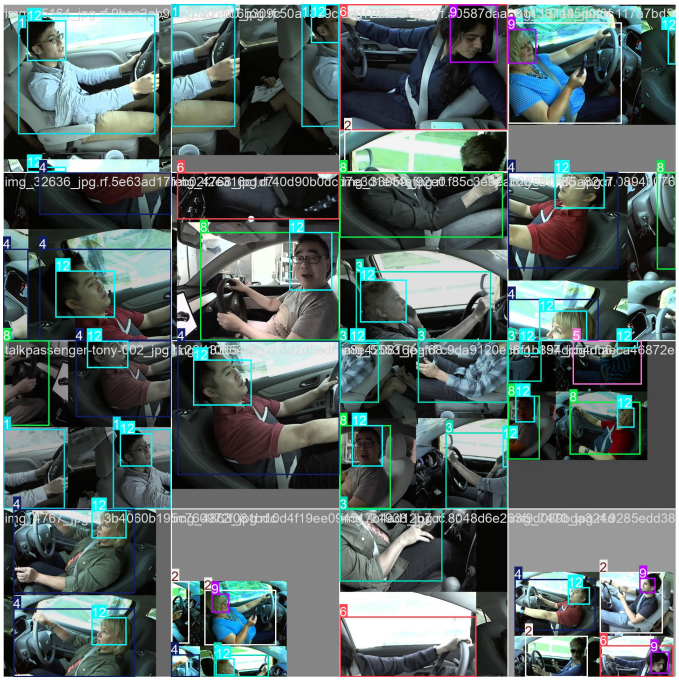


Fig. 3. Dataset samples.

that bounding box size is conditionally dependent on object location, which may reflect different spatial behaviors such as hand gestures occurring at the periphery and face-related actions closer to the center. These correlations provide valuable context for downstream model design, indicating potential biases or regularities in driver positioning and behavior manifestation. Understanding these feature interactions supports more robust model generalization and informed data augmentation strategies.

C. Training and Validation Performance Analysis

Fig. 6 presents the training performance of the fine-tuned RT-DETR-L model across 100 epochs for the distracted driving detection task. The plots include training and validation losses—namely Generalized Intersection over Union (GIoU) loss, classification loss, and L1 loss—as well as evaluation metrics such as precision, recall, and mean Average Precision (mAP) at different Intersection-over-Union (IoU) thresholds.

The training losses exhibit a consistent downward trend, indicating effective convergence of the model. Specifically, the *train/giou_loss* decreases from an initial value of 0.48 to approximately 0.18 by the final epoch, while the *val/giou_loss* stabilizes around 0.27. This steady reduction in GIoU loss suggests improving localization performance during both training and validation. Similarly, the classification loss (*train/cls_loss* and *val/cls_loss*) rapidly declines in the first 30 epochs and gradually converges, with final values around 0.3 (train) and 0.45 (validation). The small divergence between training and validation curves implies minor overfitting, which is acceptable given the complexity of the task.

The L1 loss, which measures box regression accuracy, also demonstrates smooth convergence in both training and validation curves, further confirming the model's ability to predict accurate bounding boxes. Importantly, the validation loss metrics closely follow their training counterparts, indicating good generalization to unseen data.

In terms of evaluation metrics, the RT-DETR-L model achieves excellent performance. Precision rises quickly and remains above 0.95, reflecting a low rate of false positives. Recall reaches approximately 0.90, showing the model's ability to detect a majority of relevant instances. The mAP@50 metric peaks at around 0.99, which is near perfect and confirms that the model can accurately detect and classify driver behaviors with moderate localization tolerance. More significantly, the mAP@50–95 metric, which accounts for stricter localization requirements, reaches a high value of 0.77. This result indicates that the model maintains strong performance even under challenging bounding box overlap conditions.

Overall, the training curves validate that the RT-DETR-L model has effectively learned to identify and localize various distracted driving behaviors. The absence of oscillations or divergence in the loss curves suggests stable optimization, and the high precision-recall scores confirm its suitability for real-time, safety-critical applications such as in-vehicle driver monitoring systems.

D. F1-Confidence Analysis

The F1-Confidence curve shown in Fig. 7 illustrates the classification performance of the fine-tuned RT-DETRL (Real-Time Detection Transformer Large) model on the distracted driving dataset. This visualization plots the F1-score across varying confidence thresholds for each class, providing an in-depth look at how the model's predictive certainty correlates with performance.

At a global level, the model achieves a peak macro-averaged F1-score of 0.97 at a confidence threshold of 0.626, represented by the thick blue curve. This high score reflects a well-calibrated balance between precision and recall, suggesting that the model performs optimally when predictions are accepted above this confidence threshold. Such a threshold is crucial for real-world deployment where misclassifications—either false positives or false negatives—can carry safety implications. Individual class curves offer further insight. Most classes, such as c1–Texting, c3–Operating the Radio, c4–Drinking, and d3–Eyes Open, maintain consistently high F1-scores across a broad range of thresholds. Their curves exhibit a plateau-like behavior near the top of the plot, indicating strong, stable performance and suggesting that the model is highly confident in these classifications. These behaviors likely present strong visual features or consistent spatial patterns that the model has effectively learned. In contrast, certain classes exhibit significantly lower F1-scores and more erratic curve behavior. For example, c7–Talking to Passenger and d0–Eyes Closed both show a gradual improvement in F1 as confidence increases but fail to reach the performance level of other classes. This could indicate visual ambiguity, high intra-class variability, or overlaps with other behaviors that confuse the model. Specifically, c7 likely suffers from feature similarity with actions like talking on the phone or looking toward mirrors, while d0 may be harder to detect due to its subtle or brief visual cues. At high confidence levels (above

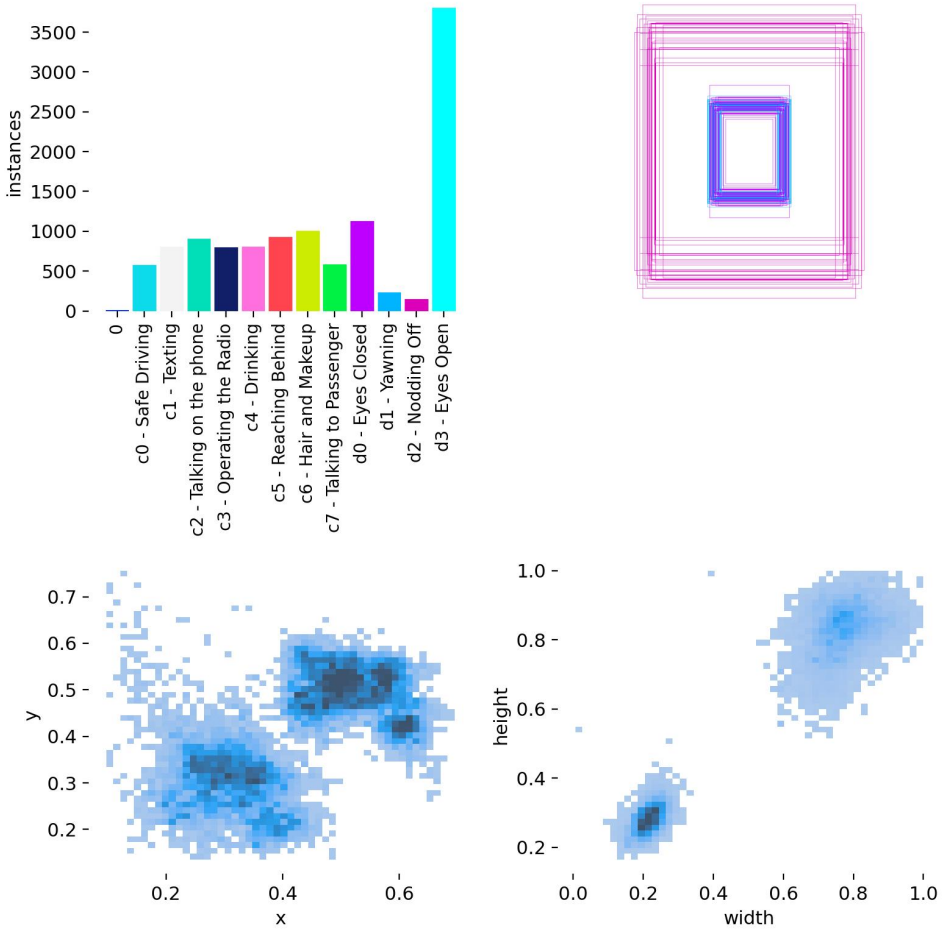


Fig. 4. Dataset analysis.

∼0.9), most curves begin to drop sharply, indicating that while precision increases, recall significantly decreases. This is expected behavior, as overly strict confidence thresholds filter out many true positives, especially for ambiguous or low-frequency classes. The drop-off underscores the importance of selecting a confidence threshold that balances both metrics effectively.

Overall, the curve suggests that RT-DETRL is well-calibrated and performs reliably across most categories, despite inherent dataset imbalances. While some classes remain challenging, the model's confidence-based F1 performance supports its suitability for real-time deployment in safety-critical systems. Further gains could potentially be achieved by applying class-specific thresholds or confidence calibration techniques such as temperature scaling.

E. Precision and Recall Analysis

Fig. 8 presents three complementary performance curves for the fine-tuned RT-DETRL (Real-Time Detection Transformer Large) model: the Precision-Confidence Curve, Recall-Confidence Curve, and the aggregated Precision-Recall Curve. Together, these plots provide a comprehensive view of how confident the model is in its predictions, and how effectively it balances the trade-offs between true positives, false positives, and false negatives across the 12 distracted driving behavior classes.

In Fig. 8a, the Precision-Confidence Curve shows how precision varies as a function of model confidence thresholds. Most classes achieve and sustain high precision values above 0.95 across a wide range of confidence levels. Notably, the macro-averaged curve (thick blue line) peaks at a perfect precision of 1.00 at a confidence threshold of 0.956. This suggests that when the model is highly confident (confidence $\geq$ 0.95), its predictions are almost always correct. However, the steep decline in precision at lower confidence thresholds highlights the importance of setting an appropriate minimum confidence during deployment to avoid introducing low-quality detections.

The Recall-Confidence Curve in Fig. 8b complements the previous analysis by demonstrating how recall degrades with increasing confidence thresholds. The macro-average recall remains high at 0.99 even at a confidence of 0.0, which is expected as low thresholds admit more predictions (increasing recall at the cost of more false positives). The recall for most classes such as c0–Safe Driving, c1–Texting, and c2–Talking on the Phone remains robust across the threshold range,

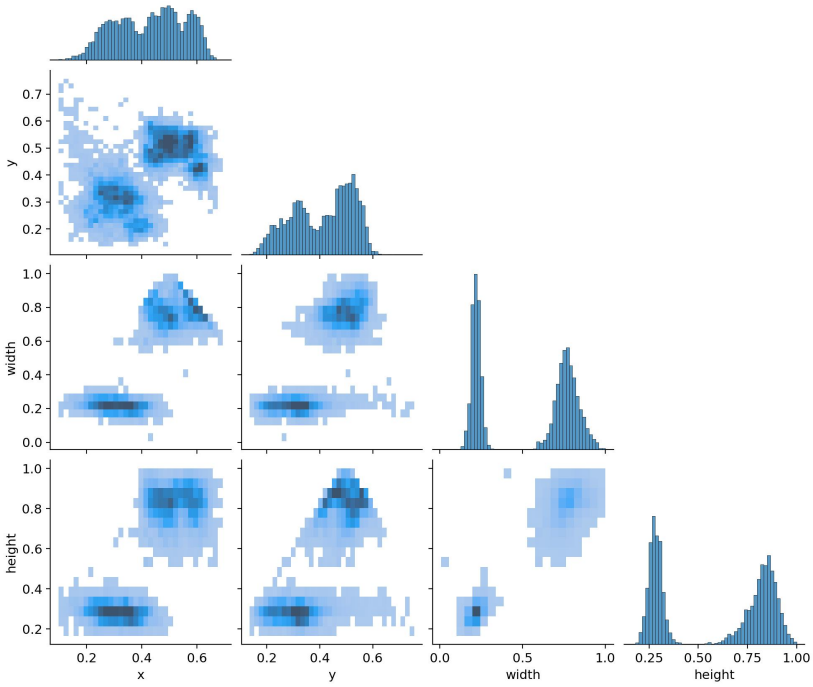


Fig. 5. Dataset correlogram.

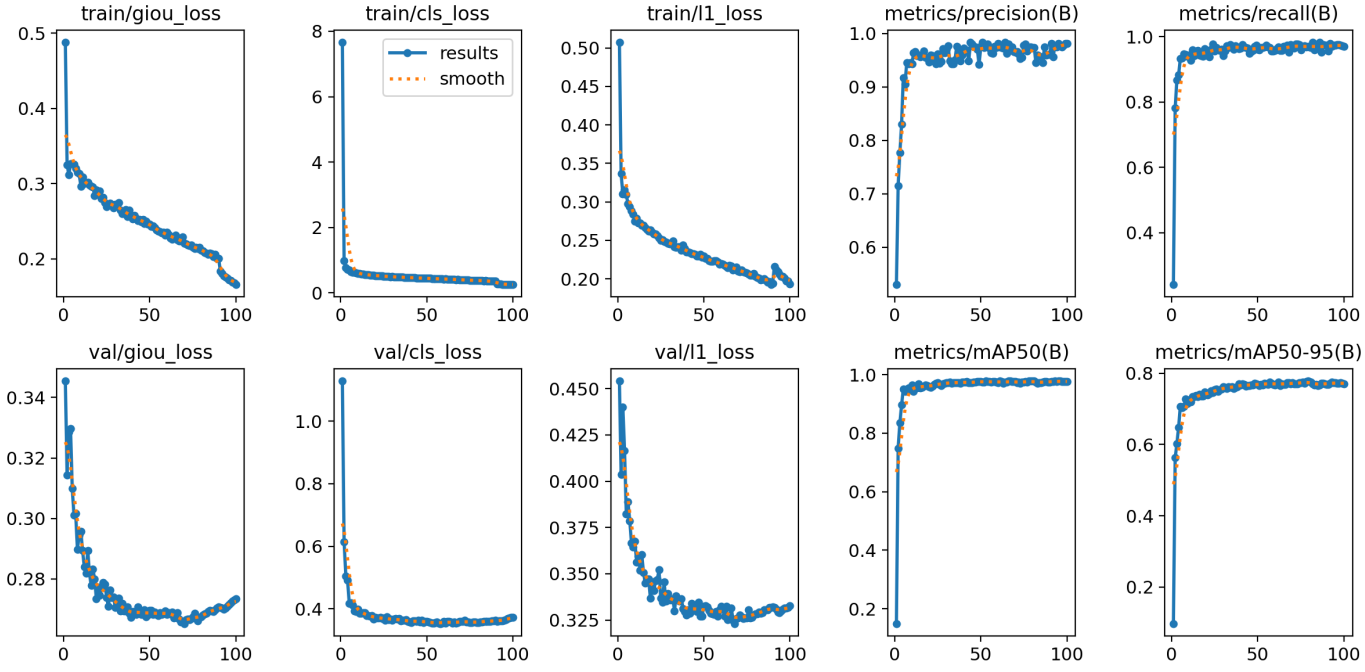


Fig. 6. Training performance of RT-DETRL.

whereas more visually subtle classes like d0–Eyes Closed show a steeper decline in recall, particularly at higher thresholds. This indicates that such classes are more prone to false negatives when confidence thresholds are stringent, potentially

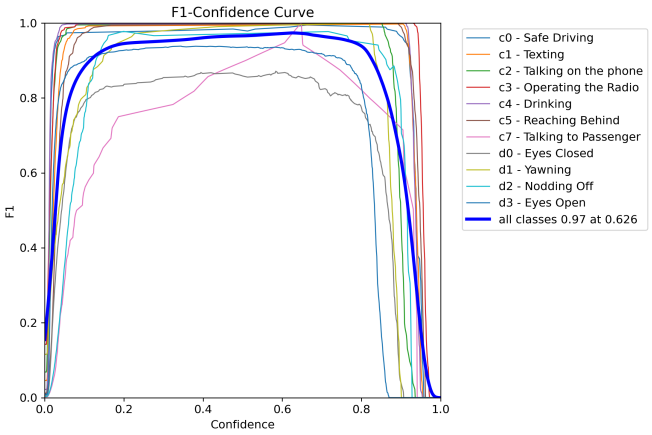


Fig. 7. F1-Score performance of fine-tuned RT-DETRL.

due to ambiguous visual features or class imbalance.

Fig. 8c shows the aggregated Precision-Recall (PR) Curve, a key diagnostic tool for evaluating classifier performance independently of any specific confidence threshold. The macro PR curve reveals strong class separability and high confidence across all behaviors, culminating in a mean Average Precision (mAP@0.5) of 0.978. Nearly all behavior classes, including c0–Safe Driving, c1–Texting, and c4–Drinking, reach near-perfect precision and recall (0.995), suggesting that the model is highly capable of detecting these behaviors accurately and consistently.

However, three classes deviate from this near-perfect trend. First, d0–Eyes Closed achieves a relatively lower recall (0.869) and exhibits a noticeable drop in both recall and precision near the right edge of the PR curve, indicating the model's difficulty in detecting this subtle and short-duration action. Similarly, d2–Nodding Off and d3–Eyes Open show minor but visible degradation, with recall values of 0.950 and 0.940 respectively. These slight declines suggest that transient or ambiguous facial expressions and head positions are more challenging for the model to disambiguate, likely due to intra-class variation and class overlap.

In conclusion, the precision and recall analysis confirms that RT-DETRL exhibits outstanding detection performance across most driver behavior classes, with a particularly high degree of confidence calibration and class separability. While some edge cases—especially those involving subtle facial cues—still present challenges, the overall mAP and class-level metrics underscore the model's robustness and readiness for real-time deployment. Fine-tuning class-specific confidence thresholds or employing hard negative mining techniques could further enhance performance for the most challenging classes.

F. Confusion Matrix Analysis

Fig. 9 presents the normalized confusion matrix for the fine-tuned RT-DETRL model on the distracted driving dataset. This matrix provides detailed insight into class-wise prediction accuracy and misclassification patterns, allowing for a granular evaluation of the model's strengths and weaknesses across the 12 driver behavior categories and the background class.

Overall, the matrix indicates strong performance on the majority of classes, particularly those with well-defined visual cues. For instance, the model correctly classifies nearly all instances of c1–Texting, c2–Talking on the Phone, c3–Operating the Radio, and c4–Drinking, with negligible misclassifications. These classes appear as strong diagonal blocks in the matrix, reflecting high confidence and distinct visual signatures associated with these behaviors. d3–Eyes Open emerges as the most frequent class, with 494 correct predictions. However, it is also subject to notable confusion with other visually similar classes. Specifically, it is misclassified as d0–Eyes Closed (42 times), d2–Nodding Off (7 times), and even as background (41 times). This indicates a limitation in distinguishing fine-grained facial states—such as open versus closed eyes—especially under challenging lighting or occlusion conditions.

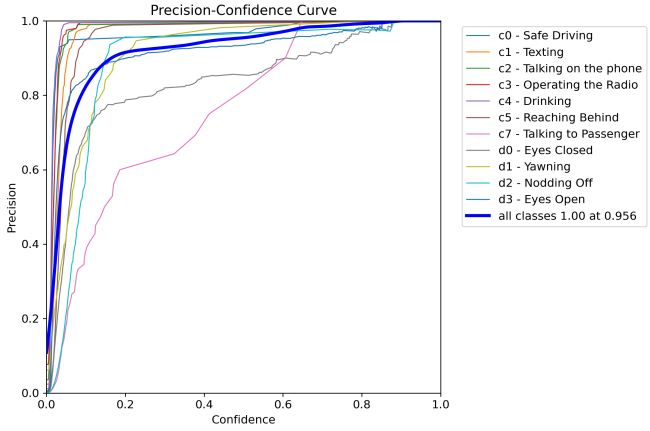
Another class with relatively strong performance is d0–Eyes Closed, correctly predicted 159 times. However, it is frequently confused with d3–Eyes Open (21 times) and background (42 times), likely due to the subtle visual differences and the temporally brief nature of this behavior. Similarly, d1–Yawning exhibits dispersed misclassifications into adjacent categories like d0–Eyes Closed (23 times) and d3–Eyes Open (14 times), suggesting intra-class variability and possible annotation overlap. Among the most ambiguous categories is c7–Talking to Passenger, which is predicted correctly only 7 times and misclassified widely across multiple classes, including d3–Eyes Open and d0–Eyes Closed. This may reflect insufficient visual differentiation, poor representation in training data, or annotation noise. Additionally, c6–Hair and Makeup appears to be entirely missed by the model, with no true positives in its row, underscoring the need for further data augmentation or class rebalancing. The background class is also a notable source of confusion, absorbing a significant number of misclassifications across multiple foreground classes. This suggests that the model occasionally struggles to separate driver behavior cues from background context, especially in edge cases or non-standard poses.

In summary, the confusion matrix analysis reinforces prior precision-recall findings and highlights that while the model performs exceptionally well on dominant and visually distinctive behaviors, challenges remain in recognizing subtle or overlapping actions, especially those involving facial expressions or body posture. Improving performance in these areas may require targeted strategies such as class-specific data augmentation, temporal modeling, or attention-based refinement modules to better isolate and interpret subtle driver cues.

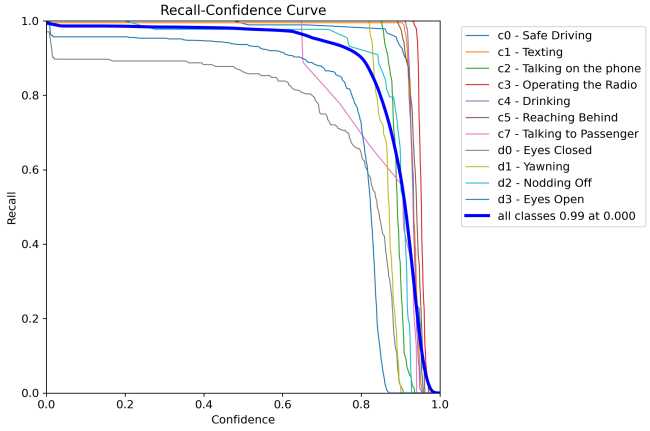
V. Detection Results

Table II presents the detection performance of RT-DETRL across the driver monitoring dataset. The table includes the number of images and instances per class, as well as the mean Average Precision at IoU thresholds 50–95 (mAP50–95). Additionally, the average inference speed, preprocessing, and post-processing time per image are reported.

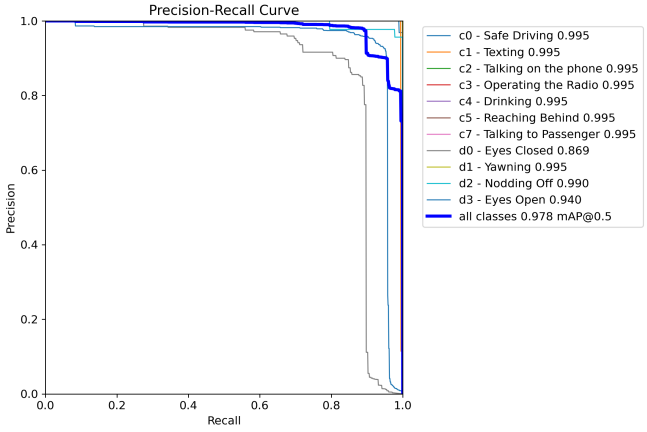
The RT-DETRL model achieves an overall mAP50–95 of 0.774 across all classes. It demonstrates strong performance in identifying well-defined behaviors such as safe driving (0.919), operating the radio (0.917), and texting (0.849). Challenging behaviors like eyes open (0.568) and yawning (0.656) show



(a) Precision curve.



(b) Recall curve.



(c) Precision-recall curve.

Fig. 8. Precision and recall for fine-tuned RT-DETRL.

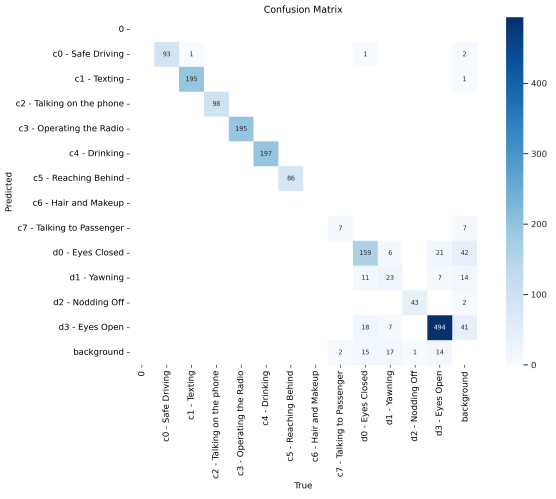


Fig. 9. Confusion matrix of fine-tuned RT-DETRL.

relatively lower precision, potentially due to visual ambiguity and overlapping features among drowsiness-related classes.

TABLE II. RT-DETRL DETECTION RESULTS

Class	Images	Instances	mAP50–95
All	1000	1711	0.774
Safe Driving	93	93	0.919
Texting	196	196	0.849
Talking on the Phone	98	98	0.792
Operating the Radio	195	195	0.917
Drinking	197	197	0.863
Reaching Behind	86	86	0.835
Talking to Passenger	9	9	0.842
Eyes Closed	204	204	0.823
Yawning	53	53	0.656
Nodding Off	44	44	0.752
Eyes Open	536	536	0.568
Speed: 0.2 ms preprocess, 8.5 ms inference, 0.3 ms postprocess per image			

Despite this, the model maintains fast inference speeds and accurate classification of most driver states.

The prediction examples shown in Fig. 10 illustrate the effectiveness of the fine-tuned RT-DETRL model in detecting various driver behaviors and states on validation data. The images present real-world in-cabin scenarios with multiple driver actions including drowsiness-related behaviors such as Eyes

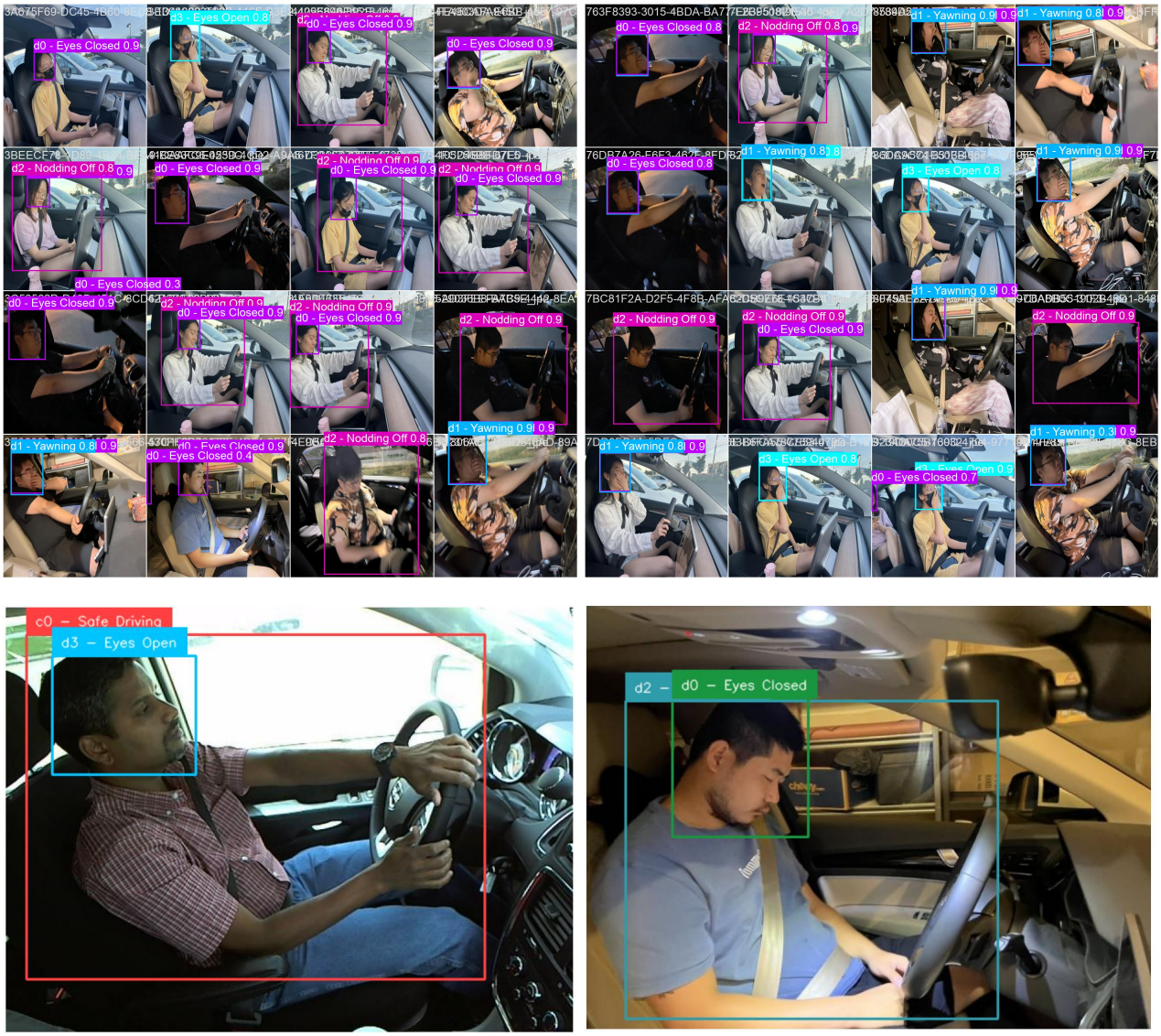


Fig. 10. Prediction examples using fine-tuned RT-DETRL on validation data.

Closed, Yawning, and Nodding Off, as well as the alert state Eyes Open. Each detection is annotated with a bounding box and a confidence score. Notably, the model demonstrates high consistency in localizing and distinguishing between visually similar states like Eyes Closed and Nodding Off, which often co-occur. The bounding boxes are accurately placed around the driver's face and upper body regions, capturing relevant features critical for behavior classification. Confidence scores are generally high (e.g., 0.9+), reflecting the model's reliability in these cases.

VI. Edge Deployment on ARM/DPU Platform

To assess the feasibility of deploying the fine-tuned RT-DETRL model on resource-constrained embedded systems, we evaluated its performance on an ARM-based edge AI platform. Specifically, we quantized the model to 8-bit integer precision (INT8) using quantization-aware training techniques and exported it for deployment on an NVIDIA Jetson Orin Nano, representative of modern ARM-based edge processors.

Table III summarizes the real-time performance metrics. The quantized model achieved a processing speed of approximately 17.5 frames per second (FPS) at a resolution of 640×640 pixels, with an average inference latency of 57 ms per frame. Despite significant reductions in memory footprint and computational cost, the quantized RT-DETRL retained a competitive accuracy, with an mAP50–95 of 0.761 compared to 0.774 from its original floating-point version. Peak RAM usage remained under 1.1 GB, and the device operated within a 9W power envelope during continuous inference.

These results demonstrate that the RT-DETRL model, when properly optimized, is highly suitable for real-time deployment on embedded platforms. This makes it a strong candidate for safety-critical applications such as in-vehicle driver monitoring systems, where reliable behavior detection and low-latency performance are essential under limited hardware resources.

TABLE III. QUANTIZED RT-DETRL PERFORMANCE ON ARM-BASED EDGE AI PLATFORM (JETSON ORIN NANO)

Metric	Value
Model Size (INT8)	27.8 MB
mAP50–95 (INT8 vs FP32)	0.761 (↓0.013)
Inference Speed	17.5 FPS
Latency per Frame	57 ms
Preprocessing / Postprocessing Time	0.4 ms / 0.5 ms
Peak RAM Usage	1.1 GB
Power Consumption	9W (avg)

VII. COMPARATIVE STUDY

Recent research in distracted driver detection, illustrated in Table. IV, demonstrates a clear evolution from handcrafted CNN classifiers toward lightweight real-time detection models and, more recently, transformer-based architectures. Early CNN-based approaches such as Drive-Net [19] achieved high accuracy (95%) by combining deep features with classical classifiers (e.g., random forests), but they lacked the ability to perform spatial localization—limiting their use to pure behavior classification. To address efficiency on constrained hardware, subsequent works such as the Decreasing Filter CNN (DFCNN) [20] refined convolutional kernel designs, achieving competitive accuracy (98.3%) while maintaining reduced computational complexity, thus enabling deployment on edge devices.

A second line of research focused on lightweight detection with enhanced feature fusion. The LCNN-MSSF model [22] exemplifies this trend by incorporating multi-scale feature fusion into a compact CNN, striking a balance between accuracy and cost for mobile platforms. Similarly, pruning-based optimization has produced highly compact models, such as P-YOLOv8 [21], which compresses YOLOv8-Tiny to under 3MB while sustaining real-time performance (18.2 FPS) and achieving over 99% accuracy on embedded hardware. These models demonstrate the feasibility of real-time inference but still remain mostly restricted to coarse behavior classification rather than spatially resolved detection.

What emerges from this literature is a trade-off between accuracy and localization: CNN-based classifiers achieve strong recognition scores but cannot provide object-level spatial insights, while pruned or lightweight detectors optimize speed and memory but often compromise contextual reasoning. Our fine-tuned RT-DETR-L model advances beyond these trade-offs by integrating both transformer-based contextual modeling and object-level detection into a unified pipeline. Unlike prior CNN or pruned detector variants, RT-DETR-L achieves superior accuracy (mAP50 = 0.995, mAP50–95 = 0.774) while simultaneously enabling fine-grained localization. Its attention-driven design captures interdependencies between driver posture and actions, addressing limitations of conventional CNN backbones.

To provide a more analytical perspective, we highlight the quantitative contrasts between the reviewed methods. Drive-Net reached 95% accuracy, whereas DFCNN improved this by 3.3% to 98.3%. LCNN-MSSF reported 97.8%, demonstrating efficiency but falling short of the best CNN baseline. The pruned P-YOLOv8 achieved 99.1% accuracy with a compact 2.8MB footprint and 18.2 FPS on Jetson hardware, making it the strongest prior work. In comparison, our RT-DETR-L achieves a mAP50 of 0.995 (99.5%), representing a 0.4% improvement over P-YOLOv8 and a 1.2% gain over DFCNN. Crucially, RT-DETR-L also reports mAP50–95 about 0.774, providing object-level localization metrics absent in CNN-based classifiers. Despite its transformer complexity, the model sustains 17.5 FPS on ARM-based edge devices, nearly matching P-YOLOv8 while offering richer localization capabilities. These quantitative comparisons underscore RT-DETR-L's balance of high accuracy, fine-grained detection, and deployment feasibility, setting it apart from prior approaches.

Therefore, the progression in distracted driver detection methods illustrates a movement toward models that are simultaneously lightweight, accurate, and deployable. Within this trajectory, RT-DETR-L represents a synthesis of these advances, combining the semantic depth of transformer architectures with the efficiency required for real-world, real-time in-cabin monitoring systems.

VIII. CONCLUSION AND FUTURE WORK

Although the proposed RT-DETR-L framework demonstrates strong potential for distracted driving detection, several promising avenues remain for future research. One direction is the incorporation of temporal modeling to capture dynamic patterns in driver behavior. Extending the current image-based approach to video sequences using transformer encoders or recurrent attention modules could allow earlier prediction of distraction and provide richer contextual understanding of posture and gaze dynamics. Similarly, combining multiple modalities such as facial landmark tracking, gaze estimation, and physiological cues like blink duration or yawning frequency would improve robustness under occlusion, low-light conditions, or camera blind spots.

Another important area is hardware optimization for deployment in real vehicles. While the model already runs efficiently on embedded devices, further compression through quantization, pruning, or neural architecture search could ensure ultra-low-latency operation on power-constrained platforms such as ARM Cortex-M microcontrollers, FPGAs, or automotive-grade NPUs. This would enable wide-scale adoption in safety-critical Advanced Driver Assistance Systems.

From a learning perspective, future studies could explore self-supervised and semi-supervised training strategies to alleviate the dependency on large annotated datasets, which are costly to acquire in real-world driving scenarios. Federated learning also presents an attractive solution for privacy-preserving adaptation across vehicle fleets, enabling broader generalization without centralized data sharing. Expanding datasets to encompass diverse drivers, vehicle types, and environmental conditions will further improve model robustness in heterogeneous contexts.

Finally, meaningful progress will require system-level integration and attention to broader societal factors. Distracted driving detection should be coupled with in-cabin and vehicle dynamics data, ensuring seamless interaction with ADAS pipelines and enabling proactive interventions such as adaptive warnings or automated control handovers. At the same time, future research must address privacy, security, and regulatory concerns, establishing transparent benchmarks and explainable

TABLE IV. COMPARISON OF RT-DETRL WITH SELECTED SOTA METHODS

Method	Architecture	Dataset / Metric	Performance	Inference Remarks
Drive-Net [19]	CNN + Random Forest	Custom dataset	95.0% accuracy	Conventional, CPU-based
CNN w/ Decreasing Filter Size [20]	CNN (DFCNN)	State Farm	98.3% accuracy	Efficient and scalable
P-YOLOv8 [21]	Pruned YOLOv8-Tiny	AUC Distracted Driver	99.1% accuracy	2.8MB model, 18.2 FPS on Jetson
LCNN with MSFF [22]	Lightweight CNN + MSFF block	Real-world dataset	97.8% accuracy	Designed for low-end devices
RT-DETRL (Ours)	Transformer (RT-DETR-L)	In-cabin behavior dataset	mAP50: 0.995; mAP50–95: 0.774	Real-time inference with high localization

Note: Accuracy is classification accuracy unless otherwise stated. mAP indicates object detection performance across IoU thresholds.

AI mechanisms to foster user trust. Together, these directions highlight the path toward practical, ethical, and scalable deployment of transformer-based distracted driving detection systems in next-generation intelligent vehicles.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number "NBU-FFR-2025-3693-01".

REFERENCES

[1] A. Mustapha, A. M. Abdul-Rani, N. Saad, and M. Mustapha, "Advancements in traffic simulation for enhanced road safety: A review," *Simulation Modelling Practice and Theory*, vol. 137, p. 103017, 2024.

[2] E. Papadimitriou, A. P. Afghari, D. Tselentis, and P. van Gelder, "Road-safety-II: Opportunities and barriers for an enhanced road safety vision," *Accident Analysis & Prevention*, vol. 174, p. 106723, 2022.

[3] A. Festag, A. Hessler, R. Baldessari, L. Le, W. Zhang, and D. Westhoff, "Vehicle-to-vehicle and road-side sensor communication for enhanced road safety," in *Proc. 15th World Congress on Intelligent Transport Systems*, Nov. 2008.

[4] C. Alparslan, M. F. Yentimur, T. Kütük-Sert, and Ş. Bayraktar, "A review on additive manufactured engineering materials for enhanced road safety and transportation applications," *Polymers*, vol. 17, no. 7, p. 877, 2025.

[5] M. Malinverno, J. Mangues-Bafalluy, C. E. Casetti, C. F. Chiasserini, M. Requena-Esteso, and J. Baranda, "An edge-based framework for enhanced road safety of connected cars," *IEEE Access*, vol. 8, pp. 58018–58031, 2020.

[6] R. Zhao, S. H. Tang, J. Shen, E. E. B. Supeni, and S. A. Rahim, "Enhancing autonomous driving safety: A robust traffic sign detection and recognition model TSD-YOLO," *Signal Processing*, vol. 225, p. 109619, 2024.

[7] M. Sabir, M. Suhail, M. Umarulla, and M. Yousuf, "Enhancing transportation safety with YOLO-based CNN autonomous vehicles," in *Proc. 2024 Int. Conf. Electronics, Computing, Communication and Control Technology (ICECCC)*, pp. 1–8, IEEE, May 2024.

[8] M. A. Al-Qaness, A. A. Abbasi, H. Fan, R. A. Ibrahim, S. H. Alsamhi, and A. Hawbani, "An improved YOLO-based road traffic monitoring system," *Computing*, vol. 103, no. 2, pp. 211–230, 2021.

[9] A. Charef, Z. Jarir, and M. Quafafou, "Enhancing road safety: Automated traffic violation detection and counting system using YOLO algorithm," in *Proc. 2024 Mediterranean Smart Cities Conf. (MSCC)*, pp. 1–6, IEEE, May 2024.

[10] J. Zhu, X. Li, P. Jin, Q. Xu, Z. Sun, and X. Song, "MME-YOLO: Multi-sensor multi-level enhanced YOLO for robust vehicle detection in traffic surveillance," *Sensors*, vol. 21, no. 1, p. 27, 2020.

[11] X. Liu, Y. Chu, Y. Hu, and N. Zhao, "Enhancing intelligent road target monitoring: A novel BGS YOLO approach based on the YOLOv8 algorithm," *IEEE Open Journal of Intelligent Transportation Systems*, 2024.

[12] Q. Shen, L. Zhang, Y. Zhang, Y. Li, S. Liu, and Y. Xu, "Distracted driving behavior detection algorithm based on lightweight StarDL-YOLO," *Electronics*, vol. 13, no. 16, p. 3216, 2024.

[13] Y. S. Poon, C. Y. Kao, Y. K. Wang, C. C. Hsiao, M. Y. Hung, Y. C. Wang, and C. P. Fan, "Driver distracted behavior detection technology with YOLO-based deep learning networks," in *Proc. 2021 IEEE Int. Symp. Product Compliance Engineering-Asia (ISPCE-Asia)*, pp. 01–05, IEEE, Nov. 2021.

[14] N. Tanaka, H. Tanaka, M. Ikeda, and L. Barolli, "A comparative study of four YOLO-based models for distracted driving detection," in *Int. Conf. Emerging Internet, Data & Web Technologies*, pp. 362–370, Cham: Springer Nature Switzerland, Feb. 2024.

[15] F. Sajid, A. R. Javed, A. Basharat, N. Kryvinska, A. Afzal, and M. Rizwan, "An efficient deep learning framework for distracted driver detection," *IEEE Access*, vol. 9, pp. 169270–169280, 2021.

[16] R. Salakapuri, N. K. Navuri, T. Vobbilineni, G. Ravi, K. Karmakonda, and K. A. Vardhan, "Integrated deep learning framework for driver distraction detection and real-time road object recognition in advanced driver assistance systems," *Scientific Reports*, vol. 15, no. 1, p. 25125, 2025.

[17] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., ... & Chen, J. (2024). Detrs beat yolos on real-time object detection. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16965-16974).

[18] Ipylot project, *Distracted Driving Dataset*, Open Source Dataset, Roboflow Universe, Roboflow, July 2022. Available: https://universe.roboflow.com/ipylot-project/distracted-driving-v2wk5 (visited on 31-Aug-2025).

[19] R. Majdi, A. Boudour, H. Messaoud, and M. Hammami, "Drive-Net: Convolutional neural network for driver distraction detection," *Procedia Computer Science*, vol. 170, pp. 1187–1192, 2020.

[20] B. Qin, J. Qian, Y. Xin, B. Liu, and Y. Dong, "Distracted driver detection based on a CNN with decreasing filter size," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6922–6933, 2021.

[21] M. R. Elshamy, H. M. Emara, M. R. Shoaib, and A. H. A. Badawy, "P-YOLOv8: Efficient and accurate real-time detection of distracted driving," in *Proc. IEEE High Performance Extreme Computing Conference (HPEC)*, pp. 1–6, 2024.

[22] Y. Li, P. Xu, Z. Zhu, X. Huang, and G. Qi, "Real-time driver distraction detection using lightweight convolution neural network with cheap multi-scale features fusion block," in *Proc. Chinese Intelligent Systems Conference*, vol. II, pp. 232–240, Springer, 2021.

Breast Cancer Classification Using Ensemble Voting: A Feature Selection Approach

Antu Kumar Guha[1], Jun-Jiat Tiang[2]*, Abdullah-Al Nahid[3]*
Electronics and Communication Engineering Discipline, Khulna University, Khulna-9208, Khulna, Bangladesh[1,3]
Centre for Wireless Technology-CoE for Intelligent Network-Faculty of Artificial Intelligence and Engineering,
Multimedia University, Persiaran Multimedia, 63100 Cyberjaya, Selangor, Malaysia[2]

***Abstract*—Breast cancer is one of the most common and deadly diseases affecting women around the worldwide. It is specially affecting in regions where has limited access to advanced diagnostic tools. Recent studies have shown that blood-based biomarkers can give a cost-effective alternative for early detection. This paper represents a machine learning-based approach for classifying breast cancer using clinical and biomedicial data. We have used the Breast Cancer Coimbra dataset for our study. We employed four filter-based feature selection methods—Mutual Information, Chi-Square, ANOVA F-test, and Pearson Correlation Coefficient—to identify the most relevant features for classification. We have applied two classifiers (AdaBoost and Ensemble Voting Classifier) to enhance predictive accuracy. The ensemble model achieved an accuracy of 82.86%. Key features such as glucose, HOMA, insulin, resistin, and age consistently contributed across all selected methods.It highlights that a few of the features has a great significance in breast cancer prediction. This study also try to investigate the reasons behind the missclassification cases. Our results show that using statistical feature selection with ensemble learning reasonable helps to boost the accuracy of breast cancer prediction. This approach helps the model focus on the most important features.**

***Keywords*—Breast cancer; machine learning; feature selection; ensemble learning; AdaBoost; biomedical data classification**

I. Introduction

Cancer is a deadly disease in which some of the body's cells grow uncontrollably and spread to other parts of the body [1]. According to World Health Organization (WHO), 10 million people died of cancer in 2020 [2]. It is nearly a sixth of all cancer deaths. There are over 100 different kinds of cancer. One of these cancers is breast cancer (BC) which is mainly affecting women.

In 2022, the world cancer research fund estimated that there were nearly 2,296,840 new BC cases [3].It was nearly 11.6 percent of all cancer cases. Moreover, almost 670 000 deaths were reported because of BC in the same year by WHO [4]. Fig. 1 illustrates the total incidence of breast cancer cases worldwide between the years 2015 and 2024. It has seen more cases over the years. In 2015, there were close to two million cases. It grew slowly every year, until it hit 2.5 million in 2024. This number remained the same between 2020 to 2022, potentially attributed to the impact of the COVID-19 pandemic on health services [5]. Then the numbers began to climb again. This increase explains the need for early detection and better awareness in order to reduce the risk of breast cancer [6]. Metastasis and late detection is one of the major factors that leads to high mortality among BC patients. Metastasis refers to the motion of cancer in the breast to other body parts of a human body. The metastasis increases the death rate. The discovery of BC at its early stage could save lives and make treatment effective. We can use various screening methods such as mammography, magnetic resonance imaging (MRI), ultrasound, and biopsy for BC detection [7].

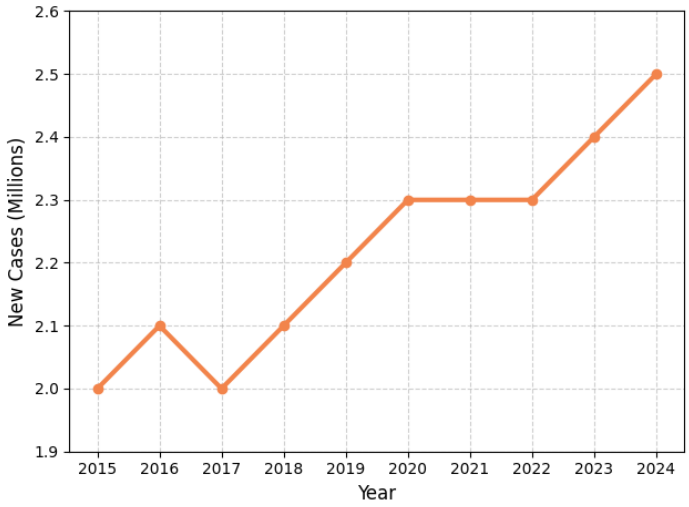


Fig. 1. Number of BC cases globally for last 10 years.

However, these methods are too expensive to afford for many women in rural area.

Also, villages lack professional physicians and adequate medical equipment [8]. The rural area mortality rate is therefore likely to be 2-4 percent higher than the urban regions. In order to address this challenge, scientists have looked at other cheaper diagnostic methods. Blood tests for specific biomarkers—like glucose, insulin, leptin, adiponectin, and others—do not directly diagnose breastcancer. But, these biomarkers can provide indicative information of metabolic health as well as inflammation within the body. This information allows to assess individual risk of disease development or rate of progression over time. The identification of these biomarkers may improve access to care as the blood tests are easy to perform in rural environments.

The Breast Cancer Coimbra (BCC) dataset is a popular dataset for the BC classification problem. This dataset consists of nine features (seven blood-based markers, age, and BMI). Glucose level is a measure of sugar in the human blood stream.

*Corresponding authors.

All cells in the body use glucose for energy. Hormones such as insulin and glucagon control glucose levels in our body. As a glucose test is also critical to the detection and monitoring of breast cancer. Unlike in normal cells, cancer cells metabolize glucose at a more rapid rate. Increased glucose utilization can be detected by means of glucose tests. Leptin is a hormone produced by adipose tissue that is involved in the regulation of appetite, energy expenditure, and metabolism more generally [9], [10]. Leptin can play a significant role in breast cancer detection and progression.

Resistin is a hormone which is mainly released immune cells in humans. It causes inflammation and helps cancer cells grow and spread [11]. High resistin levels are often found in people with cancer. Adiponectin is a hormone produced by fat cells. It is also very important to control blood sugar and reduce inflammation. Insulin is a hormone produced by the pancreas for regulating blood sugar. It allows the body to burn sugar for energy or store it for use later. Insulin also is postulated to contribute to breast cancer development at high levels because it promotes rapid growth of cancer cells [12], [13]. Checking insulin levels may help with early detection and prevention. The Homeostatic Model Assessment (HOMA) is a way to measure how resistant the body is to insulin (HOMA-IR) and how well the insulin-producing cells (beta-cells) are working (HOMA-B) [14]. Because insulin resistance and high blood sugar are linked to a higher risk of breast cancer, HOMA can help in early detection and assessing risk [14]. Monocyte Chemoattractant Protein-1 (or MCP-1) is a molecule that is very important in inflammation and immune system [15]. More recent studies show that MCP-1 is also involved in detecting breast that cancer, disease progression, and its spread (metastasis) [15].

Machine learning (ML) has shown promising performance in the prediction of early stage breast cancer in recent years, using blood based biomarkers. By using the Breast Cancer Coimbra Dataset (BCCD) Hernández-Julio et al. reached an accuracy of 95.90% using a novel approach of clustering combined with the use of pivot tables, the results were assessed through 10 folds cross validation [16]. Singh also used k-NN and achieved a remarkable accuracy of 92.11% with a 67-33 training-testing split. His research results showed to contribute to improve prediction including the use of specific features such as BMI and resistin [17]. Polat and Senturk developed a new hybrid model based on Median Absolute Deviation, feature weighting through K-means clustering and AdaBoost classifier. This method achieved 91.37% of accuracy [18]. Likewise, Akben constructed a rule based expert system implementing decision trees that achieved an accuracy of 90.52% and found important ranges of values for hormone and obesity-related variables that are linked to breast cancer [19]. Islam and Poly also utilized the k-NN algorithm with 10 folds of cross-validation and obtained an accuracy of 86.00% [20].

While earlier work achieved strong accuracy through individual classifiers or ensemble techniques of hybrid normalization. Among them few have successfully merged various statistical feature selection methods to enhance model robustness. In addition, there was low focus on applying these feature selection methodologies to an ensemble setting. To fill this gap, the present work suggests an ensemble voting classifier with four filter-based feature selection algorithms-Mutual Information, Chi-Square, ANOVA F-test, and Pearson Correlation Coefficient-added to the AdaBoost classifier. This ensures that the most significant biomarkers are used in training, improving accuracy of predictions and interpretability.

This study contributes in three main ways.

1) Methodological: We build a new AdaBoost-based ensemble that modifies a range of different filter feature-selection methods instead of changing only the classifiers. This makes the model more precise and stable.

2) Analytical: By comparing the output of different feature selectors, we find a group of major biomarkers (*Glucose, HOMA, Insulin, Resistin, and Age*) that appear repeatedly and are biologically relevant to breast cancer.

3) Applicational: Since our model uses simple filtering methods, it runs very fast and can be used even in clinics or hospitals with limited computing facilities.

Overall, this study clearly explains the new ideas and practical benefits of our approach from both the artificial intelligence and medical points of view.

Moreover, the proposed framework identifies significant advantages such as efficient computation, understandable model behavior for interpretability, and stable performance with limited data. With many light-weight statistical feature-selection methods integrated via AdaBoost, the model achieves fast training and inference without compromising interpretability. The ensemble structure also ensures stability and generalization when there are only a few medical samples. These techniques increase the uniqueness of this study and prove its utility for real-world breast cancer prediction applications.

II. Methodology

This study has used machine learning techniques to classify breast cancer cases based on clinical and biochemical data. Fig. 2 shows the overall workflow of the proposed breast cancer classification model. The process includes data preprocessing, feature selection using four statistical methods (Mutual Information, Chi-Square, ANOVA, and Correlation), AdaBoost-based classification, and ensemble voting to produce the final prediction. AdaBoost is applied due to its strong performance in binary classification problems. The feature selection techniques help identify which features are most relevant before training the models.

A. Dataset

We used the publicly available Breast Cancer Coimbra dataset from the UCI Machine Learning Repository, collected at the University Hospital Center of Coimbra (Portugal).This dataset is selected because it provides clinically interpretable, low-cost, blood-based biomarkers in support of our aim of constructing effective and affordable cancer-screening models. This dataset contains 116 instances. Among the 116 participants, 64 were diagnosed with breast cancer, and 52 were not. Each instance in the dataset includes nine features: Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, and MCP.1, along with a classification label. If the BC is presence the classification lavel is 2 otherwise 1. The features are all numerical and take from routine blood analysis and

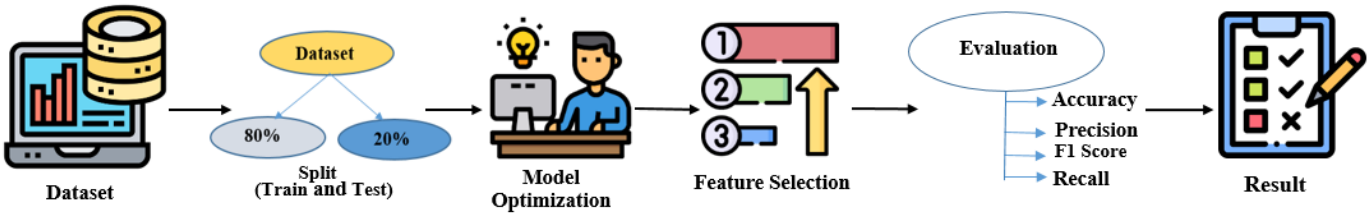


Fig. 2. The process of detecting BC.

TABLE I. FEATURES OF THE BREAST CANCER COIMBRA (BCC) DATASET

Feature	Description	Lowest Value	Highest Value	Unit
Age(f_1)	Age of the patient	24	89	Years
BMI(f_2)	Body Mass Index	18.37	37.10	kg/m^2
Glucose(f_3)	Blood glucose level	70	201	kg/dL
Insulin(f_4)	Blood insulin level	2.43	58.46	μU/mL
HOMA(f_5)	Homeostatic Model Assessment	0.50	25.05	(dimensionless)
Leptin(f_6)	Leptin hormone level	6.33	90.28	μg/mL
Adiponectin(f_7)	Adiponectin hormone level	1.65	33.75	μg/mL
Resistin(f_8)	Resistin hormone level	3.21	55.21	ng/mL
MCP-1(f_9)	Monocyte Chemoattractant Protein-1	90.09	1698.44	pg/dL

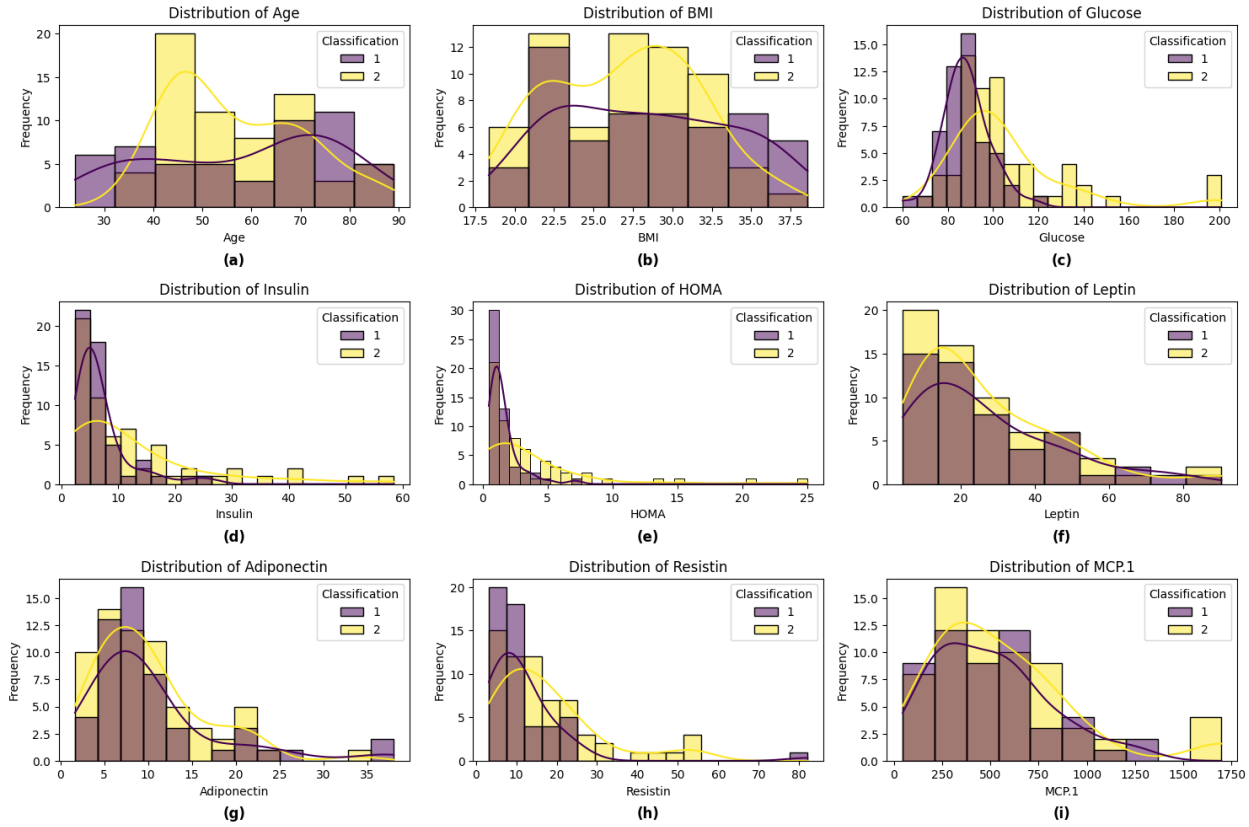


Fig. 3. Distribution of the features.

clinical measurements. According to Table I we get the highest value,lowest value and units of the features. Here,The average age of participants is about 57 years, with a minimum of 24 and a maximum of 89. BMI ranges from around 18.3 to 38.6, and glucose levels vary significantly—from 60 to 201 mg/dL, with an average of nearly 98 mg/dL. Insulin levels also show a wide range, from just 2.43 to over 58 µU/mL. Other features like HOMA, Leptin, and MCP.1 also display

noticeable variation, reflecting differences in metabolic and hormonal profiles among individuals.

Fig. 3 displays the distribution of all nine clinical and biochemical features of the Breast Cancer Coimbra (BCC) dataset, plotted separately for each class (Class 1 = no breast cancer, Class 2 = breast cancer). These histograms with kernel density estimates show the variability and overlap between features. As seen in Fig. 3 some biomarkers such as Glucose, HOMA, and Resistin exhibit evident differences between the two classes, disclosing their high potential as discriminative predictors. On the other hand, features such as Leptin and Adiponectin show significant overlap, which suggests less separability. Overall, the figure shows how statistical feature selection methods can strengthen attributes with distinctive inter-class variation, and thereby increase model accuracy and interpretability.

Before training the models, we have split the dataset into training and testing sets to evaluate the model performance. We have used various metrics such as accuracy, precision, recall, and F1-score to measure how well each classifier performs on unseen data.

B. Feature Selection

Feature selection means picking the most useful features from a dataset to help a model perform better [21].In this study, we used four common filter-based feature selection methods on the Breast Cancer Coimbra dataset. The goal was to improve the model's accuracy and remove features that are not helpful or needed.By doing this, we aim to find the most important features that help predict if a patient has breast cancer or not [22].

1) Mutual Information (MI): Mutual Information measures how much information about the output (breast cancer status) is shared with each input feature. In simple terms, it tells us how strongly a feature is related to the target. A higher MI score means the feature has a stronger dependency with the target and is more useful for prediction. We selected features with the highest MI scores.

2) Chi-square test: The Chi-Square test checks whether there is a significant relationship between a feature and the target. It works well with categorical data. Features that show a higher Chi-Square score are considered to have a stronger relationship with the target and therefore are selected.

3) ANOVA (Analysis of variance): Analysis of Variance (ANOVA) is a popular statistical method that checks if the average (mean) value of a feature is very different between two or more groups or classes [23].

Here the test statistic F-value is defined as:

$$F = \frac{\text{Variance between groups}}{\text{Variance within groups}}$$

A high F-value indicates that the feature's mean differs notably between classes. The associated p-value is used for decision-making:

- A small p-value ($p < 0.05$) suggests the feature is important for distinguishing between classes.
- A large p-value implies the feature is not useful.

In feature selection, features with low p-values and high F-values are preferred.

C. Correlation Coefficient Method

The Correlation Coefficient method measures the linear relationship between a feature and the target variable. It is usually computed using the Pearson correlation coefficient [24]:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \tag{1}$$

where x_i and y_i are individual values, and $\bar{x}$, $\bar{y}$ are their means. A higher absolute value of r indicates a stronger relationship. Features with high correlation (positive or negative) to the target are selected [25].

III. Results

For selecting the relevant features for this work we tested four statistical filtering techniques: Mutual Information, Chi-Squared Test, F-Test(ANOVA) and Correlation. All three approaches appraise feature importance from a unique statistical standpoint. Mutual Information measures how much information a feature provides about the output. From Fig. 4 according to mutual information method, Age is the most informative feature. It has the highest mutual information score. That means it may be most useful for prediction. Other features like HOMA, Resistin, and Glucose also show some importance. However, features such as BMI, Insulin, and Adiponectin have scores of zero, meaning they don't help much in this case. The chi-square test checks if there is a relationship between each characteristic and breast cancer. Similarly, from Fig. 4 we also see that according to Chi-Square test Insulin and MCP.1 have the highest scores with very low p-values, which means they are strongly related to BC. Glucose, HOMA, and Resistin are also important. On the other hand, Age and Adiponectin are not very useful according to this test. The F-Test compares the average values of features between groups. It showed that Glucose, HOMA, and Insulin are the top features. Features like Adiponectin and MCP.1 have very small F-scores. It means these features didn't vary much and not very helpful in BC detection. According to Correlation Coefficient feature selection Glucose, HOMA, and Insulin again come out on top, showing a moderate relationship with the disease. Features like Leptin and Adiponectin had values near zero. So these are not useful for prediction. Here, we have used a stepwise feature addition approach to evaluate model performance across various combinations of features. The features are added one by one according to the MI score, Chi-2 score, ANOVA f-value and correlation score. Fig. 5 shows how the model performance changes as features are added stepwise based on different selection methods. The graph clearly illustrates that adding features such as Resistin and Insulin notably improves accuracy, while including less relevant biomarkers causes a slight decline.

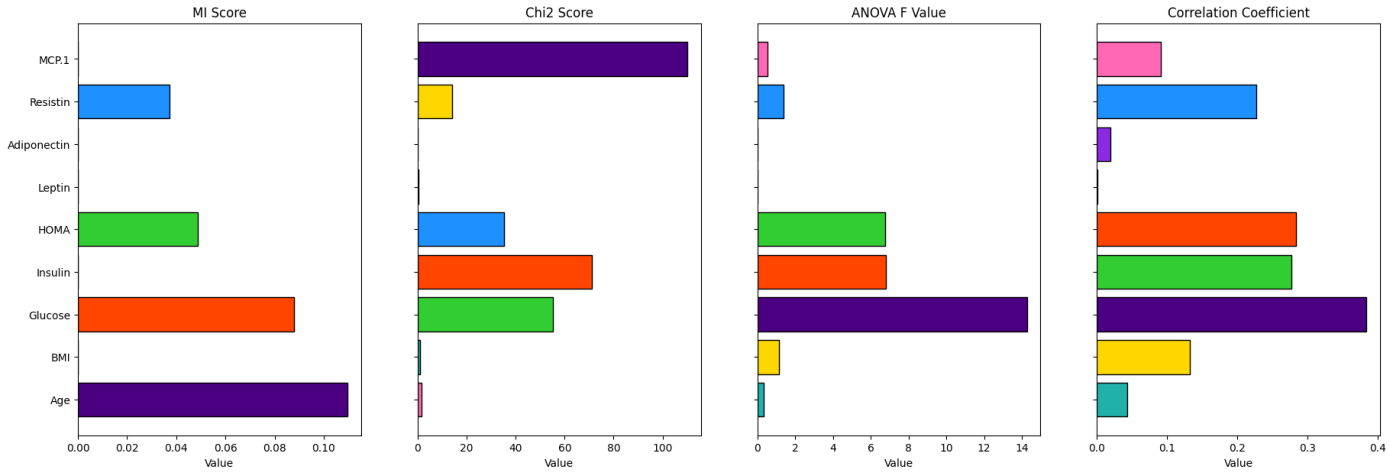


Fig. 4. Comparison of feature importance using MI score, Chi² value, ANOVA F-Statistic, and correlation coefficient.

TABLE II. FEATURE SELECTION ACCORDING TO MUTUAL INFORMATION (MI) SCORE, CHI-SQUARE SCORE, ANOVA F-VALUE, AND CORRELATION COEFFICIENT

i	**Mutual Info** ($C_{\text{mi}i}$)	**Chi-2** ($C_{\chi i}$)	**ANOVA F-value** ($C_{\text{f-value}i}$)	**Correlation** ($C_{\text{correlation}i}$)
1	$\{f_1\}$	$\{f_9\}$	$\{f_3\}$	$\{f_3\}$
2	$\{f_1, f_3\}$	$\{f_9, f_4\}$	$\{f_3, f_4\}$	$\{f_3, f_5\}$
3	$\{f_1, f_3, f_5\}$	$\{f_9, f_4, f_3\}$	$\{f_3, f_4, f_5\}$	$\{f_3, f_5, f_4\}$
4	$\{f_1, f_3, f_5, f_8\}$	$\{f_9, f_4, f_3, f_5\}$	$\{f_3, f_4, f_5, f_8\}$	$\{f_3, f_5, f_4, f_8\}$
5	$\{f_1, f_3, f_5, f_8, f_2\}$	$\{f_9, f_4, f_3, f_5, f_8\}$	$\{f_3, f_4, f_5, f_8, f_2\}$	$\{f_3, f_5, f_4, f_8, f_2\}$
6	$\{f_1, f_3, f_5, f_8, f_2, f_4\}$	$\{f_9, f_4, f_3, f_5, f_8, f_1\}$	$\{f_3, f_4, f_5, f_8, f_2, f_9\}$	$\{f_3, f_5, f_4, f_8, f_2, f_9\}$
7	$\{f_1, f_3, f_5, f_8, f_2, f_4, f_6\}$	$\{f_9, f_4, f_3, f_5, f_8, f_1, f_2\}$	$\{f_3, f_4, f_5, f_8, f_2, f_9, f_1\}$	$\{f_3, f_5, f_4, f_8, f_2, f_9, f_1\}$
8	$\{f_1, f_3, f_5, f_8, f_2, f_4, f_6, f_7\}$	$\{f_9, f_4, f_3, f_5, f_8, f_1, f_2, f_6\}$	$\{f_3, f_4, f_5, f_8, f_2, f_9, f_1, f_7\}$	$\{f_3, f_5, f_4, f_8, f_2, f_9, f_1, f_7\}$
9	$\{f_1, f_3, f_5, f_8, f_2, f_4, f_6, f_7, f_9\}$	$\{f_9, f_4, f_3, f_5, f_8, f_1, f_2, f_6, f_7\}$	$\{f_3, f_4, f_5, f_8, f_2, f_9, f_1, f_7, f_6\}$	$\{f_3, f_5, f_4, f_8, f_2, f_9, f_1, f_7, f_6\}$

TABLE III. MISCLASSIFIED INSTANCES WITH INDEX VALUES AND CORRESPONDING HEALTH METRICS

Index	Age	BMI	Glucose	Insulin	HOMA	Leptin	Adiponectin	Resistin	MCP-1	Actual Class	Predicted
26	50	38.5	106	6.7	1.7	46.6	4.6	11.7	887.1	1	2
100	74	28.6	88	3.0	0.65	31.1	7.6	18.3	572.4	2	1
31	53	36.8	101	10.1	2.5	27.1	20.0	10.2	695.7	1	2
55	34	24.2	92	21.7	4.9	16.7	21.8	12.0	481.9	2	1
114	72	25.6	82	2.8	0.57	24.96	33.75	3.27	392.4	2	1
30	66	36.2	101	15.5	3.87	74.7	7.5	22.3	864.97	1	2

1) Mutual information: The results show how adding metabolic and inflammatory markers affects the model's ability to make predictions. Initially, we have used only Age (C_{mi1}). Then the model has given moderate results with 65.7% accuracy and an F1-score of 62.5%. Interestingly, adding Glucose (C_{mi2}) and HOMA (C_{mi3}) did not improve performance. We have observed a significant gain with (C_{mi4}) when Resistin was added. It gives the accuracy to 71.4% and the F1-score to 72.2%. This means that Resistin helps improve the model's ability to differentiate between cases. The inclusion of BMI (C_{mi5}) maintained performance, while the addition of Insulin (C_{mi6}) has resulted in the highest accuracy (74.3%) and F1-score (74.3%). However, adding Leptin (C_{mi7}) and Adiponectin (C_{mi8}) slightly decreases the performance of the model. Finally, MCP-1 (C_{mi9}) recovered some of the lost performance. Then the accuracy is 71.4%.

2) Chi-square score: The results illustrate how the sequential addition of metabolic and inflammatory markers influences model performance using the Chi2 method. Initially, we used only MCP-1 ($C_{\chi 1}$), which gave an accuracy of 54.3% and an F1-score of 27.3%. This baseline performance was quite good. Then we have included Insulin ($C_{\chi 2}$). The inclusion of Insulin has decreased the performance with the accuracy of 37.1% and the F1-score to 21.4%. We have observed a significant improvement when we included Glucose ($C_{\chi 3}$). Then the accuracy jumped to 62.9% and the F1-score also reached 62.9%. This indicates that Glucose has a very important role for this classification. When we have added HOMA in ($C_{\chi 4}$), it gives a boost in the F1-score to 64.9% but accuracy remains quite similar. The addition of Resistin ($C_{\chi 5}$) further improved performance, increasing accuracy to 65.7% and the F1-score to 68.4%. The most substantial performance gain occurred when Age was included ($C_{\chi 6}$). It has increased the accuracy up to 80% and the F1-score to 78.8%. This shows that Age is a strong predictor in combination with the other markers. After adding BMI ($C_{\chi 7}$), the highest performance is maintained. The inclusion of Leptin ($C_{\chi 8}$) led to a slight decrease in

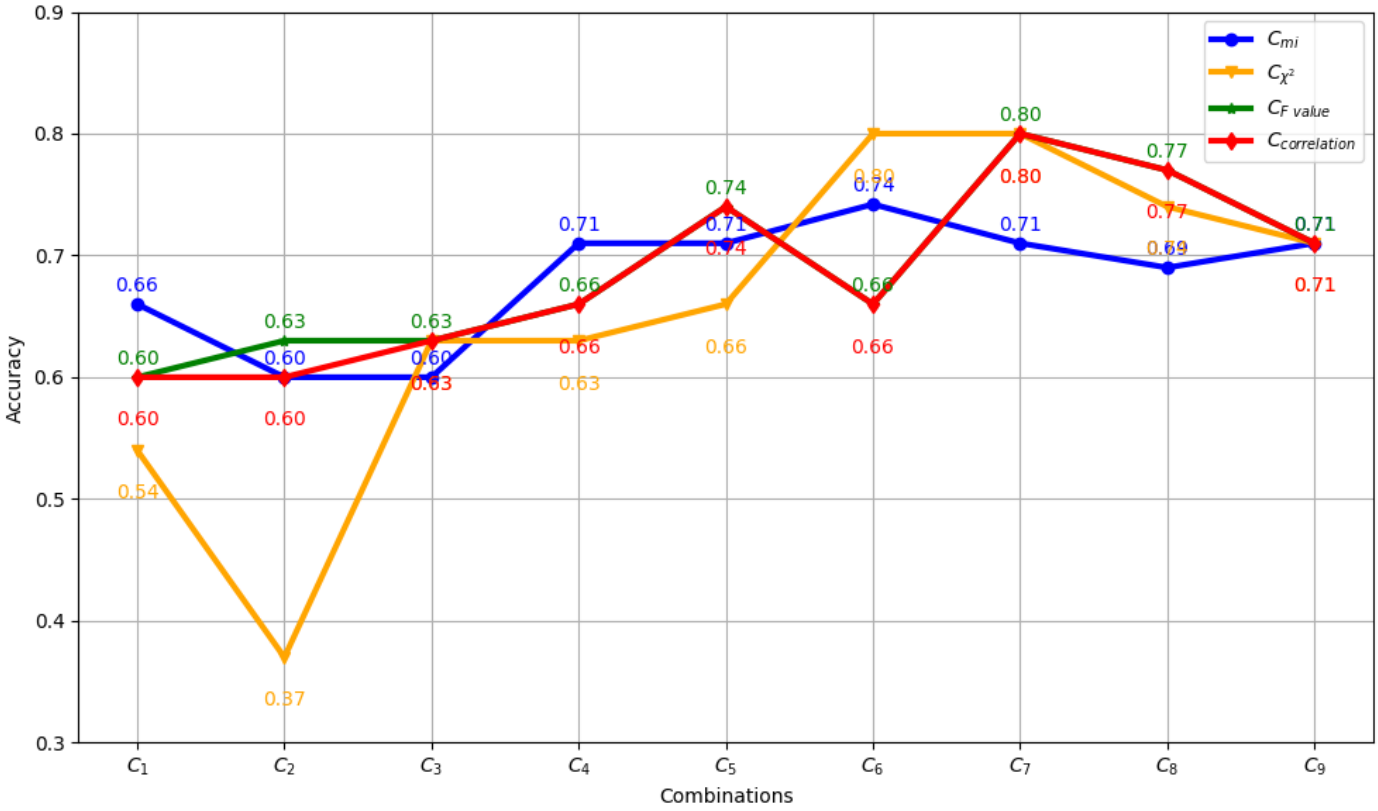


Fig. 5. Stepwise feature addition performance using mutual information score, chi-square value, ANOVA F-statistic, and correlation coefficient methods.

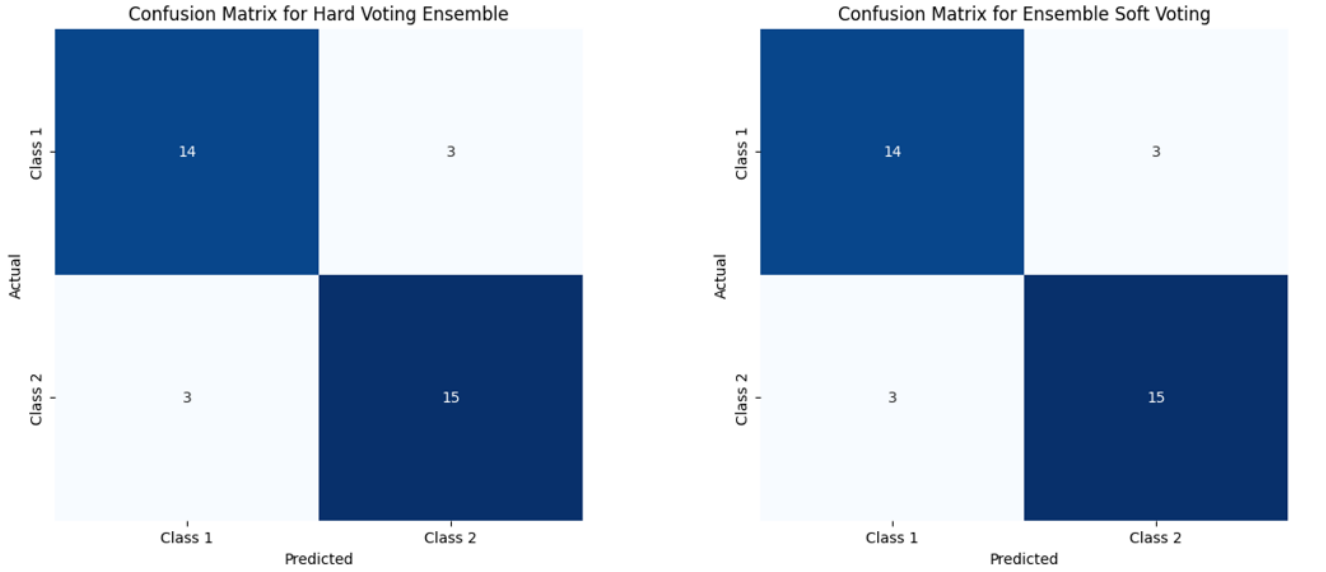


Fig. 6. Confusion matrix both for hard and soft voting.

performance, with accuracy dropping to 74.3% and the F1-score to 74.3%. Finally, adding Adiponectin ($C_{\chi 9}$) has resulted in a further decrease, with accuracy to 71.4% and the F1-score at 70.6%. This indicates that Adiponectin is not an important biomarker in this study.

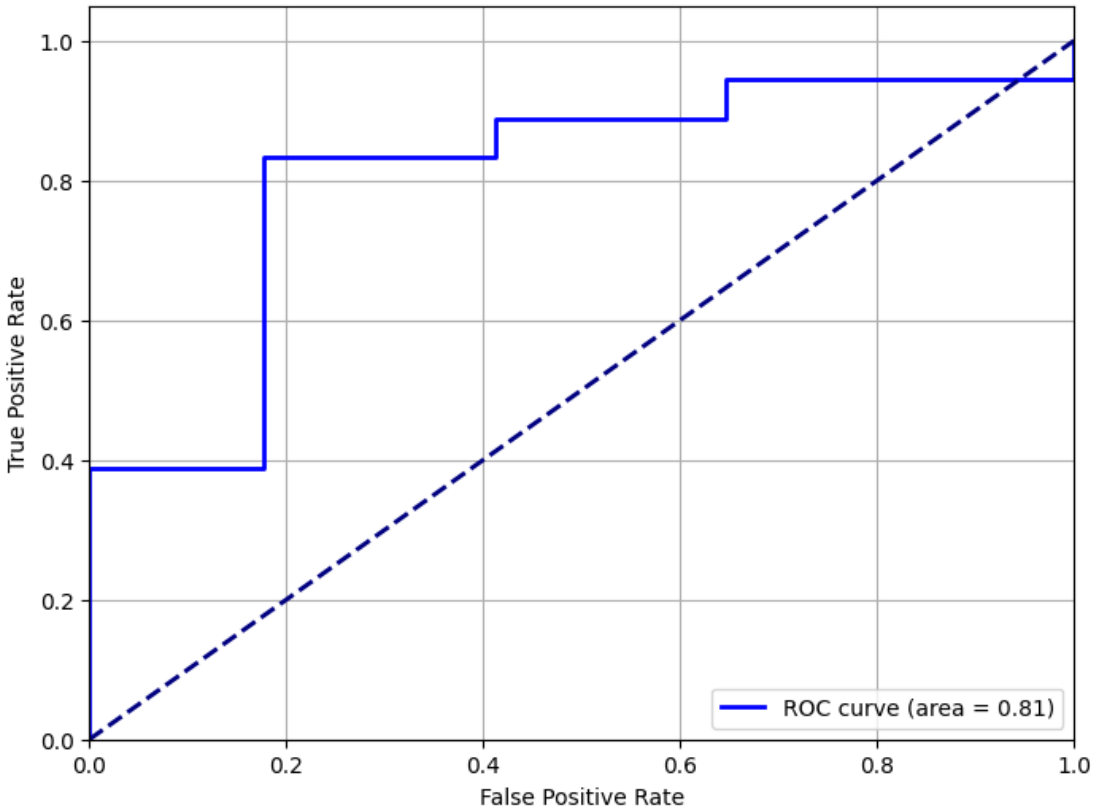


Fig. 7. Receiver Operating Characteristics (ROC) curve.

3) ANOVA f-value: The results show how progressively adding features impacts model performance using ANOVA-based selection. Initially, we have used only Glucose according to the ANOVA f-value which gives the accuracy of 60% and an F1-score of 58.8%. Then we have added Insulin ($C_{\text{f-value2}}$). It slightly improves the model performance, increasing both accuracy and F1-score to approximately 62.9%. After including HOMA ($C_{\text{f-value3}}$), it does not change the performance. However, when Resistin is added ($C_{\text{f-value4}}$), the model's performance improves to an accuracy of 65.7% and an F1-score of 68.4%. A significant boost is observed with the inclusion of BMI ($C_{\text{f-value5}}$). Then the accuracy rises to 74.3% and the F1-score to 76.9%. In this stage, the recall was 88%. When we have included MCP-1 ($C_{\text{f-value6}}$), the performance dropped. The addition of Age ($C_{\text{f-value7}}$) recovers the accuracy. Then the accuracy was found 80% and F1-score was 78.8%. When Adiponectin is added ($C_{\text{f-value8}}$), performance slightly decreases but remains robust (accuracy 77.1%, F1-score 75%). Finally, incorporating Leptin ($C_{\text{f-value9}}$) leads to a further small decline, stabilizing at 71.4% accuracy and an F1-score of 70.6%. This progression shows that some features give useful information and improve the model. But after a point, adding more features gives less benefit or can even hurt the model's accuracy.

Table II shows the ranking of features using the four selection methods. Glucose, HOMA, Insulin, Resistin, and Age appear as the most important features for breast cancer prediction.

A. Ensemble of Classifiers Trained on Distinct Feature Subsets

In this study, an ensemble voting classifier has been constructed using four pipelines. Each pipeline was based on a distinct feature selection method: mutual information, chi-square, ANOVA F-value, and correlation coefficient. The selected feature sets for these methods are as follows:

C_{mi6} = ['Age', 'Glucose', 'HOMA', 'Resistin', 'BMI', 'Insulin']

$C_{\chi 6}$ = ['MCP.1', 'Insulin', 'Glucose', 'HOMA', 'Resistin', 'Age']

$C_{\text{f-value7}}$ = ['Glucose', 'Insulin', 'HOMA', 'Resistin', 'BMI', 'MCP.1', 'Age']

$C_{\text{correlation7}}$ = ['Glucose', 'HOMA', 'Insulin', 'Resistin', 'BMI', 'MCP.1', 'Age']

Each pipeline used the AdaBoost classifier to enhance performance. The ensemble model combined predictions from the four AdaBoost-based pipelines using both hard voting and soft voting strategies. Both methods gave the same results, showing that the classifiers are well aligned. The ensemble achieved an accuracy of 82.86%, with F1-score, precision, and recall all measured at 82.35%. In Fig. 6, the confusion matrix shows that the model correctly predicted 14 instances of Class 1 and 15 instances of Class 2, with only three misclassifications for each class. This balance shows that the model performs well on both classes. The ROC curve (Fig. 7) depicts the true positive rate (TPR) versus false positive rate (FPR) across thresholds, with an AUC of 0.81 indicating fair separability.

B. Misclassified Instances

Misclassification happens when the model predicts the wrong class for a data point. In this case, the Table III shows 6 misclassified instances (indices 26, 100, 31, 55, 114, 30). These are the cases where the actual class and predicted class do not match. The main reason for misclassifications is that many features (like age, BMI, glucose) overlap between classes. This makes it hard for the model to clearly separate class 1 and class 2. Also, the model may not capture complex relationships between features. There might be some noise or variability in biological measurements.

The results section summarizes all the numerical findings, showing that the ensemble model performed consistently and well-balanced.

IV. Discussion

The experimental outcomes shows that combination of feature selection with AdaBoost-based ensemble voting significantly improves prediction accuracy. The top-rated features—Glucose, HOMA, Insulin, Resistin, and Age—are uniformly identified by all the methods. These biomarkers are of biological significance. These findings emphasize that the machine learning outcomes conform to current medical evidence, thus ensuring the validity of the suggested method.

The ensemble voting model achieved balanced accuracy for both classes, eliminating bias toward either non-cancer or cancer samples. The use of four statistical feature selection methods guaranteed that relevant and redundant variables were eliminated, increasing model interpretation and accuracy. However, some class overlaps, evidenced by feature distributions, indicate that there are biomarkers with very low discriminative power. This biological overlap explains the low number of misclassified instances indicated in the results.

Although the model is predictive, it was only based on a relatively small data set (116 cases). This restricts its use in larger populations.

It remains for future studies to advance this research by validating the model against larger, multi-center data sets and adding it to imaging or genomics to enhance diagnostic accuracy.

V. Conclusion

In this study, we applied AdaBoost and ensemble voting classifiers combined with feature selection methods to classify breast cancer cases using the Breast Cancer Coimbra dataset. By using four statistical filter techniques — Mutual Information, Chi-Square, ANOVA F-test, and Correlation Coefficient — we identified the most relevant features for prediction. Key features such as glucose, HOMA, insulin, resistin, and age were common across all selection methods. This highlights their strong predictive value and importance in breast cancer classification.

Our ensemble model reached an impressive 82.9% accuracy, with well-balanced precision, recall, and F1-scores. This shows it has strong and reliable predictive power in classifying breast cancer cases. Despite this, a few misclassifications were observed, mainly due to overlapping feature distributions and possible biological variability. These results show that combining feature selection with ensemble methods improves classification performance.

From the biomedical perspective, this study illustrates how the features glucose, HOMA, insulin, and resistin are highly correlated with breast cancer. These properties illustrate how metabolic and inflammatory problems are connected to the cancer. Because the biomarkers can be measured by simple blood tests, this procedure is a cheaper and non-invasive way to detect breast cancer at an early stage, especially where high-level medical imaging is unavailable.

From the AI perspective, this work illustrates that employing a lot of feature selection algorithms under an ensemble setting makes the system more robust and interpretable despite having little data. Both ensemble voting and AdaBoost techniques allow the model to work effectively without confusing it regarding which features to prioritize.

In the future,Utilizing the model on larger datasets from different hospitals will also make it stronger and more reliable. Overall, this study bridges medical utility and artificial intelligence and shows that interpretable ensemble learning is a useful and low-cost tool for early breast cancer detection.

Acknowledgments

This research has been supported by the Research Management Center, Multimedia University.

References

[1] National Cancer Institute, "What is cancer?" 2022, accessed: 2025-06-12. [Online]. Available: https://www.cancer.gov/about-cancer/understanding/what-is-cancer

[2] World Health Organization, "Cancer," 2020, accessed: 2024-06-01. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[3] World Cancer Research Fund, "Breast cancer statistics," 2022, accessed: 2024-06-01. [Online]. Available: https://www.wcrf.org/dietandcancer/cancer-trends/breast-cancer-statistics/

[4] World Health Organization, "Breast cancer," 2022, accessed: 2024-06-01. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[5] W. H. Organization, "Breast cancer now most common form of cancer: Who takes action," *World Health Organization*, 2021, available at: https://www.who.int/news/item/03-02-2021-breast-cancer-now-most-common-form-of-cancer-who-takes-action.

[6] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "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.

[7] National Cancer Institute, "Breast cancer screening," 2024, accessed: 2025-06-01. [Online]. Available: https://www.cancer.gov/types/breast/patient/breast-screening-pdq

[8] J. Smith and L. Brown, "Addressing the rural cancer care gap," *Journal of Rural Health*, vol. 37, no. 4, pp. 615–620, 2021.

[9] Y. Zhang, R. Proenca, M. Maffei, M. Barone, L. Leopold, and J. M. Friedman, "Positional cloning of the mouse obese gene and its human homologue," *Nature*, vol. 372, no. 6505, pp. 425–432, 1994.

[10] J. M. Friedman, "20 years of leptin: Leptin at 20: An overview," *Journal of Endocrinology*, vol. 223, no. 1, pp. T1–T8, 2014.

[11] C.-J. Lee, M. Fu, X. Liu, and Y.-C. Ko, "Resistin promotes angiogenesis in endothelial cells: involvement of the pi3k/akt pathway," *Biochemical and Biophysical Research Communications*, vol. 423, no. 1, pp. 64–69, 2012.

[12] C. J. Bailey and R. C. Turner, "Insulin and cancer risk: Epidemiological and biological perspectives," *Nature Reviews Endocrinology*, vol. 16, no. 7, pp. 393–406, 2020.

[13] D. F. Hayes, C. Isaacs, and V. Espina, "Insulin, insulin-like growth factors and breast cancer risk," *Breast Cancer Research*, vol. 18, no. 1, pp. 1–8, 2016.

[14] L. Mathews and J. Smith, "Role of homa in assessing insulin resistance and beta-cell function in breast cancer patients," *Journal of Clinical Endocrinology*, vol. 105, no. 4, pp. 1234–1241, 2020.

[15] H. Kim, S. Lee, and J. Park, "Monocyte chemoattractant protein-1 (mcp-1) and its role in breast cancer metastasis," *Cancer Immunology Research*, vol. 9, no. 7, pp. 789–797, 2021.

[16] Y. Hernández-Julio *et al.*, "Breast cancer prediction based on clustering and pivot table techniques," in *Proceedings of the International Conference (assumed)*, 2018, accuracy: 95.90%.

[17] B. K. Singh, "Determining relevant biomarkers for prediction of breast cancer using anthropometric and clinical features: A comparative investigation in machine learning paradigm," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 2, pp. 393–409, 2019.

[18] K. Polat and U. Senturk, "A novel ml approach to prediction of breast cancer: Combining of mad normalization, kmc based feature weighting and adaboostm1 classifier," in *2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Ankara, Turkey, 2018, pp. 1–5.

[19] S. B. Akben, "Determination of the blood, hormone and obesity value ranges that indicate the breast cancer, using data mining based expert system," *IRBM*, vol. 40, no. 6, pp. 355–360, 2019.

[20] M. R. Islam and Y.-T. Poly, "Breast cancer prediction using k-nn classifier on coimbra dataset," in *Assumed Proceedings*, 2020, accuracy: 86.00%.

[21] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[22] D. Dua and C. Graff, "Uci machine learning repository [breast cancer coimbra data set]," http://archive.ics.uci.edu/ml, 2017.

[23] R. A. Fisher, *Statistical Methods for Research Workers*. Oliver and Boyd, 1925.

[24] K. Pearson, "Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 187, pp. 253–318, 1896.

[25] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer, 2013.

A Two-Stage Framework for Abnormalities Detection in WCE Images by Combining Semantic Segmentation and Deformable Agent-Based Classification

Brahim Alibouch, Yasmina El Khalfaoui
LabSIV-Department of Computer Science-Faculty of Sciences, Ibn Zohr University, Agadir, Morocco

***Abstract*—Wireless capsule endoscopy (WCE) has revolutionized gastrointestinal (GI) diagnostics by offering a patient-friendly imaging and diagnostic tool compared to traditional endoscopic techniques. However, the manual assessment of these images is a time-consuming task and is prone to inaccuracies, which necessitates the implementation of automated approaches. In this paper, we introduce a two-stage deep learning framework to identify the most common GI abnormalities in WCE images. The first stage in the proposed method is to segment suspicious regions from the WCE images, which act as potential markers for GI abnormalities. In the second stage we perform frame-level classification to identify and categorize different pathologies in the GI tract. Extensive experiments conducted on four image datasets demonstrate that our approach achieves the highest values in terms of accuracy, precision, recall and specificity in comparison with four common deep learning methods : Resnet50, VGG16, Vit-S16 and InceptionV3.**

Keywords*—Wireless capsule endoscopy; deep learning; classification; gastrointestinal abnormalities

I. Introduction

Wireless Capsule Endoscopy has emerged as a revolutionary non-invasive diagnostic technology for visualizing the GI tract, particularly the small bowel, which remains challenging to access through conventional endoscopic procedures. This innovative technique involves patients swallowing a pill-sized capsule equipped with cameras that capture thousands of images as it traverses the digestive system, providing comprehensive visualization without the need for sedation or exposure to ionizing radiation [1]. The clinical significance of WCE is particularly pronounced in the diagnosis of small bowel disorders, including suspected bleeding, iron deficiency anemia, and Crohn's disease, where it serves as a first-line diagnostic tool.

However, the clinical adoption of WCE presents significant challenges that limit its full potential in routine gastroenterological practice. Each WCE examination generates an overwhelming volume of data, typically producing 50,000 to 80,000 images per patient, with only 5-10% containing pathologically relevant findings [2]. This massive data burden places substantial strain on healthcare professionals who must manually review and analyze each frame, a process that is not only time-consuming but also prone to human error and fatigue. The situation is further complicated by the fact that 20-30% of captured frames may be obstructed by bubbles, food residues, or other contaminants, making accurate diagnosis even more challenging [3].

Among the various pathological findings detectable through WCE, polyps represent a particularly important category of lesions that require precise identification and characterization. Polyps, which are abnormal tissue growths protruding from the intestinal wall, can range from benign lesions to pre-cancerous or malignant formations. Their early detection and accurate classification are crucial for determining appropriate treatment strategies and preventing progression to more serious conditions. However, polyp detection in WCE images presents unique technical challenges due to their variable size, morphology, and appearance, as well as the inherent limitations of capsule imaging, including motion artifacts, varying illumination conditions, and the presence of intestinal contents that may obscure visualization.

The integration of artificial intelligence (AI) and deep learning technologies in WCE analysis has shown tremendous promise in addressing these challenges. Recent advances in computer vision and machine learning have demonstrated the potential to reduce reading time by up to 95% while improving diagnostic accuracy [1]. Deep learning models, particularly convolutional neural networks (CNNs), have proven superior to classical machine learning methods due to their end-to-end learning capabilities and automatic feature extraction mechanisms [4]. These technologies offer the potential to not only automate the detection process but also provide objective, reproducible results that can support clinical decision making.

Despite significant progress in applying deep learning to WCE image analysis, most existing approaches focus on either classification or segmentation tasks independently. Classification methods excel at identifying the presence of pathological conditions but provide limited spatial information about lesion location and boundaries. Conversely, segmentation approaches can precisely delineate lesion boundaries but may not provide comprehensive diagnostic information about lesion characteristics and severity. This separation of tasks represents a significant limitation in clinical practice, where both spatial localization and diagnostic classification are essential for comprehensive patient care.

The development of unified framework that simultaneously perform both segmentation and classification represents a critical advancement in WCE image analysis. Such integrated approaches can leverage the complementary strengths of both

tasks, where segmentation provides precise spatial localization that can inform and improve classification accuracy, while classification results can guide and refine segmentation boundaries. This synergistic relationship is particularly important for polyp analysis, where accurate boundary delineation is essential for size estimation and morphological assessment, while classification provides crucial information about polyp type and potential malignancy risk. Furthermore, the clinical workflow in gastroenterology naturally requires both spatial and categorical information about detected lesions. Clinicians need to know not only whether a polyp is present (classification) but also its exact location, size, and morphological characteristics (segmentation). An integrated framework that provides both types of information simultaneously would significantly enhance clinical utility and diagnostic confidence while streamlining the review process.

This work presents a novel deep learning framework designed to address the dual challenges of GI abnormalities segmentation and classification in WCE images through an integrated approach that leverages the synergistic relationship between these complementary tasks. Our framework aims to provide comprehensive diagnostic support that meets the practical needs of clinical gastroenterology while addressing the technical challenges inherent in WCE image analysis.

This paper is organized as follows. In the next section the related works are reviewed. In Section III, we explain the proposed method. Section IV presents the outcomes of the experiments conducted to evaluate our approach. Finally, Section V delineates the conclusions of this paper.

II. Related Works

The application of deep learning techniques to wireless capsule endoscopy has experienced rapid growth over the past decade, with researchers exploring various approaches to automate the detection, classification, and segmentation of GI abnormalities. The foundation of automated WCE analysis lies in the development of robust classification systems capable of distinguishing between normal and pathological findings. Soffer et al. [4] conducted a systematic review and meta-analysis of deep learning applications in WCE, revealing exceptional performance across various pathological conditions. Their analysis demonstrated that CNNs achieved high pooled sensitivity and specificity scores, with particularly impressive results for bleeding detection (sensitivity 0.98, specificity 0.99) and ulcer identification (sensitivity 0.95, specificity 0.94). The superior performance of CNNs over classical machine learning methods was attributed to their end-to-end learning capabilities and automatic feature extraction mechanisms. Recent advances in CNN architectures have shown remarkable improvements in classification accuracy. Habe et al. [5] provided a comprehensive review of deep learning performance across 30 studies from 2013-2022, revealing that Xception models achieved the highest accuracy (99%) on datasets containing 53,555 images, followed by DenseNet121 with 98.17% accuracy on only 1,000 images, demonstrating the effectiveness of modern architectures in achieving high performance even with limited training data. Transfer learning approaches, particularly those utilizing ImageNet pre-trained models, have consistently shown significant performance improvements across different architectures, with Inception-ResNet-v2 achieving 98.1% accuracy on 400,000 images.

The integration of attention mechanisms has further enhanced classification performance. Bandopadhyay et al. [6] demonstrated the superiority of Vision Transformer (ViT) architectures over traditional CNNs for GI ulcer classification in WCE images. Their ViT model achieved perfect scores across all key metrics (Accuracy, Precision, Recall, and F1-Score) when compared against prominent deep learning models including ResNet50, VGG19, Inception V3, and Inception-ResNet152V2, highlighting the potential of transformer-based architectures in medical image analysis. Several researchers have explored sophisticated approaches to address specific challenges in WCE classification. Garbaz et al. [7] proposed a dual-branch deep convolutional neural network combining DenseNet121 with residual blocks specifically designed for detecting small, indistinct vascular lesions such as angiodysplasia and bleeding zones. Their architecture achieved accuracies of 99.50% and 96.96% on MICCAI 2017 and Kvasir-Capsule datasets respectively, with the residual block component contributing 12.0% and 8.78% accuracy improvements. Li et al. [8] introduced a dynamic multi-task learning framework that addresses fine-grained recognition and class imbalance challenges through the combination of semi-hard triplet loss with weighted cross-entropy loss. Their approach achieved 96.47% F1-score on Kvasir-Capsule (10-class classification) and 96.75% F1-score on CAD-CAP (3-class classification), demonstrating the effectiveness of multi-objective optimization in handling complex classification scenarios. The challenge of limited labeled data has been addressed through various innovative approaches. Muruganantham and Balakrishnan [9] proposed ACT-WISE, an uncertainty-driven active learning framework integrated with deep semi-supervised learning. Their teacher-student consistency regularization methodology achieved 97% classification accuracy and 0.95 AUC on the Kvasir-Capsule dataset, outperforming traditional semi-supervised methods like Mean Teacher (95% accuracy) while significantly reducing annotation requirements.

Recognizing the inherent temporal nature of WCE videos, several researchers have explored spatiotemporal analysis approaches. Bordbar et al. [10] pioneered the application of three-dimensional CNNs for multiclass WCE classification, processing 16 consecutive frames to leverage temporal relationships that gastroenterologists naturally consider during diagnosis. Their 3D-CNN approach achieved significant improvements over 2D methods, with 98.92% sensitivity vs. 98.05%, 99.50% specificity vs. 86.94%, and 99.20% accuracy vs. 92.60%, demonstrating the value of incorporating temporal dependencies. Prakash and Krishnamurthy [11] developed an innovative transfer learning approach combining depth estimation with evolutionary optimization for optimal keyframe selection from WCE videos. Their TUMNetv2 architecture integrated transformer encoders with UNet+ and MobileNetv2, achieving 93% accuracy, 90% precision, 85% recall, and 87% F1-score while significantly reducing computational burden through intelligent frame selection. The combination of deep features with traditional hand-crafted features has shown promising results in WCE analysis. Amiri et al. [12] proposed a hybrid approach combining deep features from pre-trained ResNet50 with hand-crafted color, shape, and texture features. Their method utilized the Expectation Maximization algorithm for

region of interest extraction and achieved 97.82% accuracy in detecting various abnormalities including ulcers, bleeding, angiodysplasia, lymphoid hyperplasia, and polyps. Caroppo et al. [13] developed a transfer learning-based expert system that combined features from ResNet50 and Inception V4 architectures. Their hybrid approach, utilizing minimum Redundancy Maximum Relevance feature selection and SVM classification, achieved 98.09% accuracy on KID Dataset 2 and 94.48% accuracy on MICCAI 2017 dataset, with strong cross-dataset generalization capabilities (91.46% and 93.28% accuracy).

While most WCE research has focused on classification tasks, segmentation approaches have gained increasing attention for their ability to provide precise spatial localization of abnormalities. He et al. [14] proposed MVANet, a Multi-view Attention Network designed to address critical challenges in endoscopic image segmentation, including boundary blurring, low contrast, and small lesion detection. Their architecture achieved state-of-the-art performance with mDice values of 93.1% and 94.9%, and mIoU values of 88.4% and 90.7% on Kvasir-SEG and CVC-ClinicDB datasets respectively. Charfi et al. [15] presented a novel approach using Dense-UNet for segmentation followed by a modified residual attention network for classification. Their method focused on detecting and segmenting red lesions, ulcers, and polyps, addressing the challenge of processing thousands of low-quality images in WCE videos. The quality of WCE images presents significant challenges for automated analysis systems. Sadeghi et al. [3] addressed this issue by creating the first publicly available multicentre dataset for clear and contaminated region segmentation in WCE images. Their dataset included pixel-level annotations for 17,593 images, categorizing contamination levels from 1-10 based on percentage of contaminated area, with contamination ranging from minimal (0.28% ± 1.29%) to nearly complete coverage (95% ± 0.0%), highlighting the critical need for robust preprocessing and quality assessment mechanisms in WCE analysis systems. The detection of rare or previously unseen pathologies represents a significant challenge in WCE analysis. Quindós et al. [16] addressed this through a novel patch-based self-supervised approach for out-of-distribution detection. Their three-stage methodology achieved AUROC scores ¿0.6 for detecting unseen pathologies including lymphangiectasia (0.767), foreign bodies (0.729), ulcers (0.795), and blood (0.677), demonstrating the potential of unsupervised approaches for identifying novel pathological conditions. Accurate localization within the GI tract is crucial for clinical interpretation of findings. Fan et al. [17] developed an automatic locating system for identifying the pylorus and ileocecal valve using ResNet-50-based classification combined with a novel sliding pin algorithm. Their approach achieved mean locating errors of 26 frames for the pylorus and 65 frames for the ileocecal valve, representing significant improvements over previous state-of-the-art methods. Object detection frameworks have been adapted for WCE analysis to provide both classification and localization capabilities. Xiao and Feng [2] developed an improved YOLOv3-based system for small intestinal lesion detection, introducing a novel labeling strategy that expanded lesion annotation areas to 1.33 times the original size. Their approach achieved 93.5% mAP with 4% improvement over standard YOLOv3, demonstrating better speed-accuracy balance while successfully identifying lesions from small to large sizes. Fernandes et al. [18] developed a deep learning-based methodology to automatically identify physician-annotated frames within WCE videos, with MobileNet achieving superior performance with 94% accuracy for first match identification and 98% accuracy for top-20 matches.

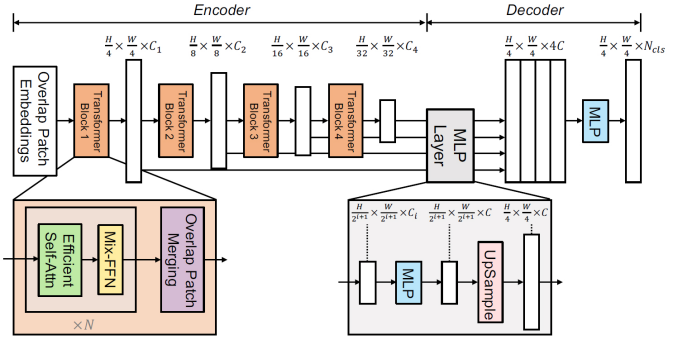


Fig. 1. SegFormer [19] architecture.

Despite significant progress in WCE image analysis, several limitations persist in current research. Most existing approaches focus on either classification or segmentation tasks independently, limiting their clinical utility where both spatial and categorical information are required. Additionally, while many studies report high accuracy on specific datasets, there remains a lack of comprehensive frameworks that address multiple aspects of polyp analysis simultaneously. The challenge of class imbalance, particularly evident in polyp detection where pathological cases are relatively rare, requires more sophisticated approaches that can effectively learn from limited positive examples while maintaining high sensitivity. Furthermore, the integration of clinical workflow considerations and real-time processing requirements has received limited attention in current research. The temporal nature of WCE videos presents both opportunities and challenges that have been insufficiently explored, and there remains significant potential for developing more sophisticated approaches that leverage the sequential nature of WCE data for improved diagnostic accuracy. Current research also lacks comprehensive evaluation frameworks that consider both technical performance metrics and clinical utility measures, making the development of standardized evaluation protocols that reflect real-world clinical scenarios essential for advancing the field toward practical clinical deployment. This comprehensive review highlights the significant progress made in applying deep learning to WCE analysis while identifying key gaps and opportunities for advancement. The proposed framework aims to address these limitations through an integrated approach that combines the strengths of both segmentation and classification methodologies for comprehensive polyp analysis in WCE images.

III. Methodology

In this paper, we introduce a two-stage deep learning framework to identify the most common GI abnormalities in WCE images. The first stage in the proposed method is to segment suspicious regions from the WCE images, which act as potential markers for GI abnormalities. To this end, we employed SegFormer-B0, a lightweight transformer-based semantic segmentation model. In the second stage we

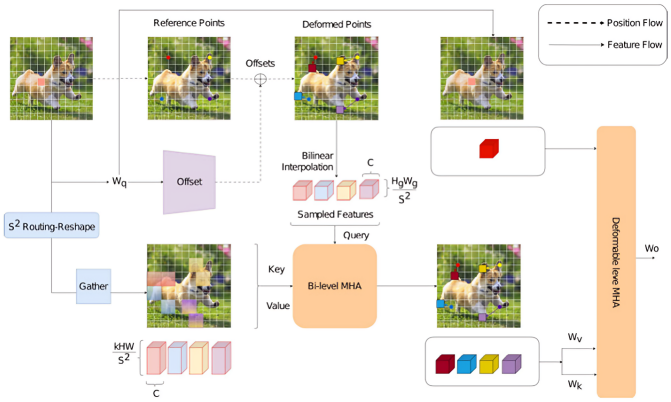


Fig. 2. Architecture of DBRA [20].

perform frame-level classification to identify and categorize different pathologies in the GI tract. For this purpose, we used DeBiFormer, a vision transformer architecture that uses an advanced attention mechanism called Deformable Agent Bi-level Routing Attention, which merges deformable attention with bi-level routing to reduce computational cost while improving accuracy.

The methodology combines the computational efficiency of SegFormer for initial segmentation with the advanced attention mechanisms of DeBiFormer for refined classification and abnormality detection.

A. SegFormer

Introduced by Xie et al. [19], SegFormer represents a significant step forward in applying Vision Transformers (ViTs) to semantic segmentation. The design philosophy of SegFormer focuses on creating a framework that is simple, efficient, and powerful, effectively overcoming the inherent limitations of prior ViT-based models in dense prediction tasks. It specifically addresses two major challenges: the low-resolution, fixed-size feature maps generated by standard ViTs, which are poorly suited for pixel-level classification, and the high computational complexity associated with self-attention, especially when dealing with high-resolution images.

The architecture consists of two main components: a hierarchical Mix Transformer (MiT) encoder and a lightweight All-MLP decoder as shown in Fig. 1. This design eliminates the need for positional encodings, supports multi-scale feature extraction, and significantly reduces model complexity while achieving state-of-the-art performance.

The encoder generates multi-scale features at resolutions of $\{1/4, 1/8, 1/16, 1/32\}$ of the original image, addressing the limitation of Vision Transformers that produce only single-scale representation and enabling the capture of both local and global features effectively, which is essential to handle dense prediction tasks like segmentation. The decoder takes the multi-scale feature maps from all four encoder stages and uses MLPs to transform and concatenate them together for pixel-wise predictions . The final output is a segmentation mask at $\frac{H}{4} \times \frac{W}{4} \times N_{cls}$ resolution, where Ncls represents the number of categories.

B. DeBiFormer

DeBiFormer (Deformable Bi-level Routing Attention Transformer) represents a significant advancement in vision transformer architectures by addressing key limitations in existing attention mechanisms. Introduced by Long et al. [20], DeBiFormer combines the flexibility of deformable attention with the semantic relevance of bi-level routing (Fig. 2) to create a more effective and interpretable attention mechanism for visual recognition tasks.

As shown in Fig. 3(a), the architecture of DeBiFormer follows a four-stage pyramid structure where each stage begins with either overlapped patch embedding (first stage) or patch merging modules (subsequent stages) to progressively reduce spatial resolution while increasing channel dimensions, followed by multiple consecutive DeBiFormer blocks. Each DeBiFormer block incorporates a 3×3 depthwise convolution for implicit relative position encoding, the DBRA module for cross-location relation modeling, and a 2-ConvFFN module for per-location feature embedding.

The core contribution of DeBiFormer is its novel Deformable Bi-level Routing Attention (DBRA) module [Fig. 3 (b)], which operates through an attention-in-attention architecture. Unlike traditional attention mechanisms that either focus on fixed patterns or lack semantic awareness, DBRA achieves both spatial flexibility and semantic relevance through a two-stage process:

- Deformable Point Generation: The module first generates deformable points using an offset network that shifts reference points toward informative regions based on query features. This allows the model to

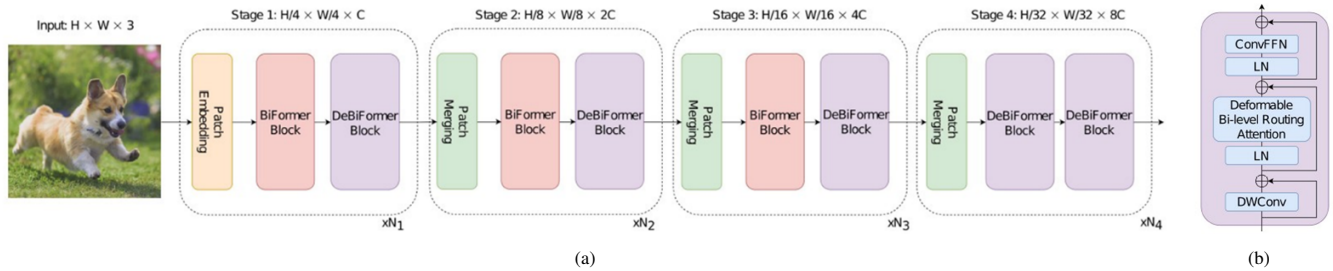


Fig. 3. (a) Overall DeBiFormer architecture, (b) DeBiFormer block [20].

adaptively focus on important areas rather than being constrained to regular grid patterns.

- Semantic Region Selection: Instead of attending to all deformed points equally, DBRA employs a bi-level routing mechanism where each deformable point acts as an agent query. These agents select only the top-k most semantically relevant regions, ensuring computational efficiency while maintaining semantic focus.

C. Proposed Method

This paper proposes a novel two-stage deep learning framework for the automated detection of abnormalities in WCE images. The overarching goal is to first precisely localize and segment potential pathological regions and subsequently classify them with high accuracy to reduce false positives. The proposed architecture, leverages the strengths of a state-of-the-art semantic segmentation model followed by a powerful, computationally efficient transformer-based classifier.

The first stage of our proposed framework is dedicated to the accurate segmentation of potential abnormalities within WCE images. For this task, we have selected SegFormer-B0, a lightweight vision transformer-based segmentation model that can effectively handle the hierarchical feature extraction required for medical image segmentation without the over-fitting issues common in traditional CNNs on small medical datasets. It processes the full WCE image and outputs a binary segmentation mask that precisely delineates the boundaries of suspicious regions. This stage acts as an intelligent attention mechanism, effectively filtering out vast areas of normal mucosa and focusing subsequent computational resources only on regions of clinical interest.

Following the segmentation of potential GI abnormalities by SegFormer-B0, the second stage of our framework involves classifying these segmented regions into specific pathological categories. For this classification task, we propose the use of DeBiFormer, leveraging its deformable agent bi-level routing attention mechanism to process both the original WCE images and the segmentation outputs from Stage 1. This dual-input approach enables the network to validate and refine segmentation results while providing robust abnormality classification. An overview of the proposed two-stage deep learning framework is illustrated in Fig. 4.

IV. Experiments

A. Datasets

For this work we used two datasets to pre-train the SegFormer-B0 model : CVC-ClinicDB [21] and Kvasir-SEG [22]. CVC-ClinicDB is a dataset consisting of 612 images extracted from colonoscopy video sequences, each with a corresponding binary ground truth mask highlighting polyps. All images are standardized at a resolution of 384 × 288 pixels, making the dataset compact and consistent. Kvasir-SEG, on the other hand, contains 1,000 images with varying resolutions, reaching up to 1920 × 1072 pixels, and each is paired with a high-quality expert-verified mask.

Once the SegFormer-B0 model pre-trained, two other datasets were used to train and test the whole proposed model. The first is Curated Colon Disease Dataset for Deep Learning [23] which includes 6000 WCE images of the GI tract, distributed across four categories: normal, ulcerative colitis, polyps, and esophagitis, with 1500 images per class. The second one is Kvasir v2 [24], a multi-class GI endoscopy dataset containing 8,000 endoscopic images evenly distributed across eight classes. Each class has 1,000 images, and the resolutions vary from 720 × 576 to 1920 × 1072 pixels.

Tables I and II provide detailed informations on each dataset.

TABLE I. Curated Colon Disease and Kvasir v2 Datasets Specifications

Dataset	Class	Train	Validation	Test	Total
Curated Colon Disease	Normal	1200	150	150	1500
	Ulcer	1200	150	150	1500
	Polyps	1200	150	150	1500
	Esophagitis	1200	150	150	1500
Kvasir v2	dyed-lifted-polyps	800	100	100	1000
	dyed-resection-margins	800	100	100	1000
	Esophagitis	800	100	100	1000
	normal-cecum	800	100	100	1000
	normal-pylorus	800	100	100	1000
	normal-z-line	800	100	100	1000
	polyps	800	100	100	1000
	ulcerative-colotis	800	100	100	1000

TABLE II. Kvasir-SEG and CVC-ClinicDB Datasets Specifications

Dataset	Train	Validation	Test	Total
Kvasir-SEG	800	100	100	1000
CVC-ClinicDB	490	61	61	612

B. Preprocessing and Data Augmentation

A combination of data augmentation and normalization techniques can improve the quality, diversity, and generalizability of training data [25]. In this work, the images are first resized to fixed size and several transformations are applied including rotations, vertical and horizontal flips, and affine transformations. Color jittering is performed to alter saturation contrast, hue and brightness also as random perspective distortion to enhance furthermore diversity. To maintain consistency during evaluation, only resizing and normalization are performed on validation set.

This preprocessing pipeline guarantees that the model encounters diverse and realistic image variations during training, while simultaneously maintaining the validation data in a standardized format for precise performance assessment.

C. Implementation Details

The SegFormer-B0 model was pre-trained with CVC-ClinicDB and Kvasir-SEG datasets using Adam optimizer and binary cross entropy loss function with a batch size of 4 over 60 epochs. The initial learning rate was set to 0.00005 and the process were executed on Google Colab, with an NVIDIA Tesla T4 GPU with 2560 CUDA cores and 16 GB of GDDR6 memory.

to train the whole model, we used Curated Colon Disease and Kvasir v2 datasets over 80 epochs using Adam optimizer, the binary cross entropy loss function, a batch size of 32

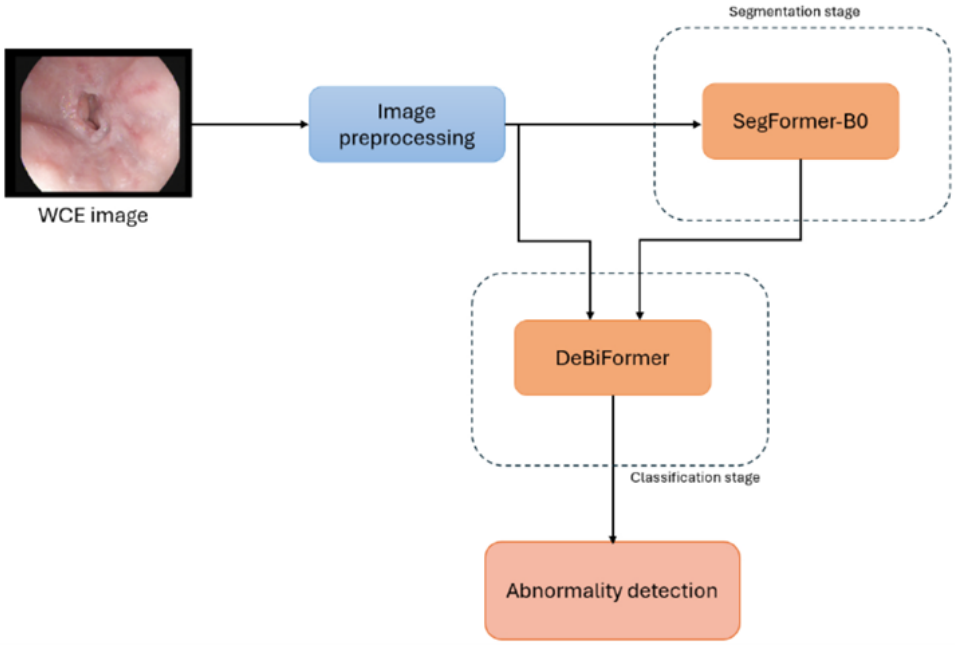


Fig. 4. Overview of the proposed method.

and an initial learning rate $\alpha = 0.0002$. The proposed model was implemented in PyTorch and trained using an NVIDIA RTX A4000 GPU with 16 GB of memory. The computational resources were provisioned through the Paperspace Gradient platform. We utilized a pre-configured GPU-optimized container, which included PyTorch 1.12, CUDA 11.6, and cuDNN 8.

D. Results

In this section, the potential of the proposed method is shown. For this purpose, comprehensive experiments are conducted based on confusion matrix analysis, which collectively offer insights into the model's performances. A confusion matrix is a tabular representation where the model's predicted class labels are compared with the ground truth labels. The matrix is structured to provide counts of four fundamental outcomes for each class:

- True Positive (TP): The model correctly predicted the positive class. (Reality: Yes, Prediction: Yes)
- True Negative (TN): The model correctly predicted the negative class. (Reality: No, Prediction: No)
- False Positive (FP): The model incorrectly predicted the positive class when it was actually negative. Also known as a Type I error. (Reality: No, Prediction: Yes)
- False Negative (FN): The model incorrectly predicted the negative class when it was actually positive. Also known as a Type II error. (Reality: Yes, Prediction: No)

Based on these basic metrics, a total of four specific performance metrics were used : accuracy, precision, specificity and recall.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}} \quad (1)$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (2)$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \quad (3)$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (4)$$

After conducting preliminary pre-training, the SegFormer-B0 model attains accuracy values of 98.80% and 91.19%, and loss values of 0.03 and 0.14 on the training and validation sets, respectively. Thus, SegFormer-B0 shows a high potential in improving the performance of segmentation in GI abnormalities detection as can be seen in Fig. 5.

In our experiments, we also compared the proposed method with four deep learning models including Resnet50 [26], VGG16 [27], InceptionV3 [28] and Vit-S16 [29]. To guarantee a scientific and logical comparison of methods, all experiments were carried out in the same environment. The comparison results are shown in Tables III and IV.

On the Curated Colon Disease dataset, our methods achieves the highest accuracy of 99.30%, followed by VGG16 (93.06%) and Vit-S16 (91.83%). Our method attains also the highest values across all other metrics reaching 99.30% precision, 99.30% recall and 99.77% specificity.

The proposed method surpasses all the four methods also On the Kvasir v2 dataset, achieving an accuracy of 94.13%

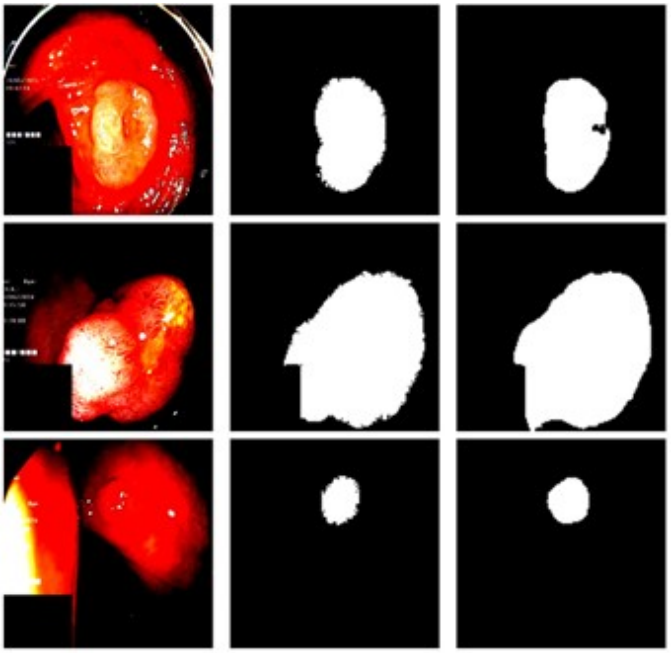


Fig. 5. Visual results of segmentation using SegFormer-B0 model. From left to right: Input image, Ground truth, SegFormer-B0 model.

TABLE III. QUANTITATIVE COMPARISON BETWEEN OUR METHOD AND DIFFERENT MODELS ON THE CURATED COLON DISEASE DATASET

Methods	Metrics			
	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
Resnet50	83.40	84.01	83	94.40
VGG16	93.06	93	90.47	97.69
InceptionV3	74.09	83.46	62.22	91.36
Vit-S16	91.83	91.98	91.73	97.33
Our method	**99.30**	**99.30**	**99.30**	**99.77**

accuracy, 94.18% precision, 94.12% recall and 99.16% specificity.

Overall, compared against four prominent deep learning models, our model achieves strong performance on both Curated Colon Disease and Kvasir v2 datasets, underscoring the efficacy and the potential of our approach.

V. CONCLUSION

In this work, we developed and evaluated a novel, two-stage deep learning framework for the automated segmentation, and classification of abnormalities in Wireless Capsule Endoscopy images. The methodology leverages the complementary strengths of two state-of-the-art Transformer-based architectures: SegFormer-B0 for precise pixel-level segmentation of potential pathological regions, and DeBiFormer for the accurate classification of these segmented regions into specific abnormality types by means of its flexible Deformable Bi-level Routing Attention module. Segmentation provides precise spatial localization that can inform and improve classification accuracy, while classification results can guide and refine segmentation boundaries. The experiments shows that our unified framework that simultaneously perform both segmentation and classification significantly enhances performances in WCE image analysis. A comparative studies was conducted to validate the effectiveness of our approach. The results show that our approach achieves the highest values in terms of accuracy, precision, recall and specificity outperforming four common deep learning methods : Resnet50, VGG16, Vit-S16 and InceptionV3.

TABLE IV. QUANTITATIVE COMPARISON BETWEEN OUR METHOD AND DIFFERENT MODELS ON THE KVASIR V2 DATASET

Methods	Metrics			
	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
Resnet50	83.40	84.01	0.83	94.40
VGG16	93.06	93	90.47	97.69
InceptionV3	74.09	83.46	62.22	91.36
Vit-S16	91.83	91.98	91.73	97.33
Our method	**94.13**	**94.18**	**94.12**	**99.16**

While our approach generally succeeds in identifying the most important anomalies, it is less efficient in the presence of less prominent anomalies and noise such as air bubbles and food debris. In future work, we aim to address these limitations and improve our model performance by including transfer Learning techniques and extending our evaluation to other public datasets offering additional variability in WCE images.

REFERENCES

[1] P. Collins, "Video-capsule endoscopy – current and future role," *Medicine*, vol. 52, no. 5, pp. 317–319, May 2024.

[2] Z. Xiao and L. N. Feng, "A Study on Wireless Capsule Endoscopy for Small Intestinal Lesions Detection Based on Deep Learning Target Detection," *IEEE Access*, vol. 8, pp. 159 017–159 026, 2020.

[3] V. Sadeghi, Y. Sanahmadi, M. Behdad, A. Vard, M. Sharifi, A. Raeisi, M. Nikkhah, and A. Mehridehnavi, "Pixel-wise annotation for clear and contaminated regions segmentation in wireless capsule endoscopy images: A multicentre database," *Data in Brief*, vol. 57, p. 110927, Dec. 2024.

[4] S. Soffer, E. Klang, O. Shimon, N. Nachmias, R. Eliakim, S. Ben-Horin, U. Kopylov, and Y. Barash, "Deep learning for wireless capsule endoscopy: A systematic review and meta-analysis," *Gastrointestinal Endoscopy*, vol. 92, no. 4, pp. 831–839.e8, Oct. 2020.

[5] T. T. Habe, K. Haataja, and P. Toivanen, "Review of Deep Learning Performance in Wireless Capsule Endoscopy Images for GI Disease Classification," *F1000Research*, vol. 13, p. 201, 2024.

[6] S. Bandopadhyay, J. R. Chowdhury, R. Shaw, S. Paul, and S. Banerjee, "Vision Transformer-Based Classification of Gastrointestinal Ulcers Using WCE Images," in *Computational Intelligence in Pattern Recognition*, A. K. Das, J. Nayak, B. Naik, M. Himabindu, S. Vimal, and D. Pelusi, Eds. Singapore: Springer Nature, 2025, pp. 367–378.

[7] A. Garbaz, S. Lafraxo, S. Charfi, M. El Ansari, L. Koutti, and M. Salihoun, "Combined deep convolutional neural networks for abnormality classification in wireless capsule endoscopy images," *Multimedia Tools and Applications*, Mar. 2025.

[8] X. Li, Q. Wu, and K. Wu, "Wireless capsule endoscopy anomaly classification via dynamic multi-task learning," *Biomedical Signal Processing and Control*, vol. 100, p. 107081, Feb. 2025.

[9] P. K. Muruganantham and S. M. Balakrishnan, "Uncertainty-driven active learning in a deep semi-supervised framework for WCE image classification," *Results in Engineering*, vol. 27, p. 106174, Sep. 2025.

[10] M. Bordbar, M. S. Helfroush, H. Danyali, and F. Ejtehadi, "Wireless capsule endoscopy multiclass classification using three-dimensional deep convolutional neural network model," *BioMedical Engineering OnLine*, vol. 22, no. 1, p. 124, Dec. 2023.

[11] M. Madhura Prakash and G. N. Krishnamurthy, "An innovative transfer learning-polyp detection from wireless capsule endoscopy videos with optimal key frame selection and depth estimation," *Biomedical Signal Processing and Control*, vol. 108, p. 107963, Oct. 2025.

[12] Z. Amiri, H. Hassanpour, and A. Beghdadi, "Combining deep features and hand-crafted features for abnormality detection in WCE images," *Multimedia Tools and Applications*, vol. 83, no. 2, pp. 5837–5870, Jan. 2024.

[13] A. Caroppo, P. Siciliano, and A. Leone, "An expert system for lesion detection in wireless capsule endoscopy using transfer learning," *Procedia Computer Science*, vol. 219, pp. 1136–1144, Jan. 2023.

[14] D. He, R. Zhang, Y. Liang, J. Chen, and Y. Li, "Multi-View Attention Network With Iterative Feature Refinement and Boundary Awareness for Endoscopic Image Segmentation," *IEEE Access*, vol. 13, pp. 131 217–131 233, 2025.

[15] S. Charfi, M. E. Ansari, L. Koutti, A. Ellahyani, and I. Eljaafari, "Modified residual attention network for abnormalities segmentation and detection in WCE images," *Soft Computing*, vol. 28, no. 9, pp. 6923–6936, May 2024.

[16] A. Quindós, P. Laiz, J. Vitrià, and S. Seguí, "Self-supervised out-of-distribution detection in wireless capsule endoscopy images," *Artificial Intelligence in Medicine*, vol. 143, p. 102606, Sep. 2023.

[17] S. Fan, S. Wei, K. Wei, S. Liu, Y. Fan, B. Lv, and L. Li, "A novel automatic locating method for pylorus and ileocecal valve in wireless capsule endoscopy," *Biomedical Signal Processing and Control*, vol. 100, p. 106969, Feb. 2025.

[18] R. Fernandes, A. Pessoa, J. Nogueira, A. Paiva, I. Paçal, M. Salgado, and A. Cunha, "Evaluation of Deep Learning Models in Search by Example using Capsule Endoscopy Images," *Procedia Computer Science*, vol. 239, pp. 2065–2073, Jan. 2024.

[19] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers," Oct. 2021.

[20] N. H. B. Long, C. Zhang, Y. Shi, T. Hirakawa, T. Yamashita, T. Matsui, and H. Fujiyoshi, "DeBiFormer: Vision Transformer with Deformable Agent Bi-level Routing Attention," Oct. 2024.

[21] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, vol. 43, pp. 99–111, Jul. 2015.

[22] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-SEG: A Segmented Polyp Dataset," Nov. 2019.

[23] F. J. P. Montalbo, "Diagnosing gastrointestinal diseases from endoscopy images through a multi-fused CNN with auxiliary layers, alpha dropouts, and a fusion residual block," *Biomedical Signal Processing and Control*, vol. 76, p. 103683, Jul. 2022.

[24] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt, M. Riegler, and P. Halvorsen, "KVASIR: A Multi-Class Image Dataset for Computer Aided Gastrointestinal Disease Detection," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, ser. MMSys'17. New York, NY, USA: Association for Computing Machinery, Jun. 2017, pp. 164–169.

[25] Y. E. Khalfaoui, B. Alibouch, and A. F. E. Ouafdi, "Combining Deep CNNs and Superpixels for Polyp Segmentation in Colonoscopy Images," in *2023 10th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, Oct. 2023, pp. 1–6.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Dec. 2015.

[27] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," Apr. 2015.

[28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," Dec. 2015.

[29] 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 16x16 Words: Transformers for Image Recognition at Scale," Jun. 2021.

Predicting Employee Attrition in the Saudi Private Sector Using Machine Learning

Haya Alqahtani, Hana Almagrabi, Amal Alharbi
Department of Information Systems-Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—Employee attrition represents a prominent issue facing organizations, as human capital represents one of the most valuable resources. Attrition refers to the voluntary or involuntary reduction in the number of employees, which can negatively affect profitability, reputation, and overall organizational performance. Therefore, a comprehensive understanding of this phenomenon, its causal factors, and the development of retention strategies is crucial for mitigating employee turnover. The purpose of this work is to predict employee attrition in the Saudi private sector and identify the key factors contributing to employee turnover using machine learning approaches. in addition, the research structurally evaluates the performance of multiple Machine Learning (ML) algorithms within the proposed framework to determine the most effective predictive model for employee attrition. This study utilized a training dataset obtained from an online survey targeting employees in the Saudi private sector in order to investigate employee attrition and identify its most prominent causes within this context. Thus, various Machine Learning (ML) algorithms, including Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), Bagging ensemble, and Voting Classifier (VC) were evaluated. The results demonstrate that the Voting Classifier (VC) yielded the highest accuracy at 90%. Moreover, the analysis identified job opportunities and job titles as some of the most influential factors driving employee turnover.**

***Keywords*—Employee attrition; attrition prediction; predictive models; machine learning; voting classifier; ensemble methods; Saudi private sector; employee turnover; employee retention; feature importance**

I. Introduction

Employee attrition is a critical challenge for organizations due to its impact on productivity, operational costs, and long-term business sustainability. Attrition occurs when employees leave their organizations due to resignation, retirement, or involuntary termination. There are various key factors influencing employee attrition, including job satisfaction, work–life balance, compensation, work environment, and age. Employees who perceive themselves as undervalued or undercompensated are more likely to leave their organizations in pursuit of better opportunities [1]. Employee attrition can be measured using metrics such as sales growth, return on equity, customer service quality, and profitability [2].

As a matter of fact, voluntary turnover is particularly disruptive, resulting in the loss of institutional knowledge and interrupting ongoing projects [3]. Moreover, new hires require significant time to achieve the performance level of their predecessors, which reduces organizational efficiency [4], [5]. Studies have shown that voluntary employee turnover creates significant costs due to the expenses of recruitment, hiring, training, and employee development, which can be more than 1.5 times an employee's annual salary [6].

Given the significant impact of attrition, predictive analytics has become an essential tool for forecasting employee turnover and formulating strategies to mitigate it [7]. Identifying employees prone to attrition enables organizations to proactively improve retention, enhance employee satisfaction, and reduce turnover-related costs [8]. HR analytics, known as talent analytics [9], plays a vital role in quantifying workforce factors that influence company outcomes, thereby supporting better decision-making [10]. Consequently, HR professionals must develop an in-depth understanding of employee profiles and the critical factors influencing turnover to accurately identify individuals at risk of departure [11], [12], [13].

Data is a crucial element for the success of a people analytics team [14]. However, in HR analytics, the focus should shift from simply collecting large volumes of data to ensuring the data provides meaningful insights and is used effectively to create organizational value [10]. Previous research on attrition prediction has predominantly utilized publicly available datasets, such as the IBM Employee Attrition dataset, as well as datasets obtained from platforms like Kaggle. Nevertheless, such datasets do not capture the complexity of real-world employee behavior and organizational dynamics, in addition to being characteristically imbalanced, causing predictive bias if not properly handled [8]. In order to address existing limitations, the primary objective of this study was to develop a realistic dataset and propose a robust ML framework for predicting employee attrition.

The main contributions of this paper are threefold:

1) Development of a contextually-grounded dataset for the Saudi private sector: This research introduces an empirically-collected dataset from 1,191 real employees across diverse Saudi industries, addressing the critical limitation of prior studies that rely predominantly on synthetic datasets (e.g., IBM HR dataset). The dataset uniquely captures region-specific factors shaped by Saudization policies and cultural dynamics, while incorporating previously underexplored variables such as job opportunities, emotional commitment, health issues, and sector diversity. This contribution fills a significant gap in attrition research for emerging markets and provides a reusable resource for future studies in similar contexts [15].

2) A comprehensive comparative ML framework with systematic feature selection: This study systematically evaluates eight ML algorithms—including conventional methods (LR, SVM, KNN, DT) and ensemble techniques (RF, XGBoost,

Bagging, Voting Classifier)—across seven feature selection configurations with SMOTE-based class balancing [16]. The Voting Classifier with RFE-selected features achieves 90% accuracy, demonstrating that ensemble methods significantly outperform conventional approaches while maintaining interpretability. This design provides generalizable methodological guidance for predictive modeling in real organizational settings where data imbalance and feature redundancy are prevalent. This framework provides HR practitioners with evidence-based guidance for selecting optimal modeling strategies for imbalanced attrition datasets.

3) Actionable insights revealing a paradigm shift in attrition drivers: The feature importance analysis reveals that dynamic developmental factors (job opportunities, career advancement, job titles) significantly outweigh traditional demographic factors (age, Academic degree) in predicting attrition. The study provides data-driven visions for targeted interventions—including career development programs and advancement pathways—that enable organizations to proactively identify at-risk employees.

Experimental results demonstrate the effectiveness of the proposed approach, with promising accuracy and generalization capabilities of the models. The remainder of the paper is structured as follows: Section II presents a review of previous studies on employee attrition prediction. Section III details the methodology adopted in this study, including dataset development, modeling framework, and evaluation strategy. Section IV reports the experimental results and feature analysis. Section V discusses the findings and their implications. Section VI offers concluding remarks and potential directions for future research.

II. Literature Review

A. Employee Attrition Overview

Employee attrition is defined as the departure of employees from an organization for various reasons, including resignation, retirement, or death [17]. Attrition is also described as a reduction in the workforce resulting from these departures and is sometimes referred to as total turnover or wastage [18], [19]. While attrition reflects a decline in the number of employees, employee turnover specifically denotes the replacement of departing employees with new hires. These terms are often used interchangeably in workforce analyses due to their conceptual overlap. Turnover can be classified as voluntary or involuntary [20]. Even though dysfunctional turnover, characterized by the loss of high-performing employees, is detrimental to an organization, the turnover of employees with easily replaceable skills tends to have less impact [21].

In fact, employee attrition is a critical issue impacting various sectors, particularly those where employee roles are essential, creating barriers to profitability, competitiveness, and productivity by causing overtime, delayed project delivery, customer dissatisfaction, and reduced morale among remaining staff who must fill vacancies [19]. Apart from this, high attrition rates contribute to increased employee turnover, resulting in substantial expenses related to recruitment, training, and development of new hires [20], [22]. Therefore, retaining employees is financially crucial as it directly influences revenue growth through customer acquisition and retention [23]. Furthermore, voluntary turnover results in workforce movement across companies and roles, transferring valuable knowledge to competitors, and consequently posing a major risk to organizational interests [24].

B. Factors Influencing Employee Attrition

While human resources represent one of the most valuable assets contributing to a company's success, one of the most prominent challenges faced by HR management is employee attrition [22]. The initial step towards addressing this issue involves identifying the underlying causes and examining the factors associated with attrition. In this sub-section, several studies that have identified or investigated these factors are discussed.

For instance, the study in [17] reportes that employee attrition occurs for various reasons, including insufficient support for employees in lower organizational positions, lack of direct communication lines between these employees and their supervisors, monotonous tasks leading to a loss of creativity in repetitive jobs, limited career advancement prospects, and a lack of opportunities. As well as that, authors in [25] confirm that employees tend to leave their jobs when they perceive inadequate compensation, experience disagreements with their managers, or confront issues related to work hours, career development, and family obligations.

Moreover, a study conducted on 40 private financial institutions—including banks, investment fund organizations, and insurance companies—in the Kottayam district of India concluded that a strained relationship between management and employees, employee dissatisfaction, lack of job security, and job-related pressure significantly contribute to employee attrition [19]. Additionally, extensive research highlights a range of factors influencing turnover rates, which can broadly be categorized into organizational factors, job-related factors, work environment factors, and personal factors [26].

C. Employee Attrition and Retention in Saudi Private Sector

In this subsection, studies investigating the factors that impact employee turnover in the Saudi private sector are presented to highlight the significance of this issue and the necessity for continued research. The following studies explore some organizational and job-related factors closely related to employee attrition, such as job satisfaction, work participation, work pressures, rewards, effective training, and recognition.

For example, the study in [27] examined the impact of training and development opportunities following Saudization on employees' intentions to leave their organizations. The findings indicated that providing employee training contributed to retention exceeding five years, with training and development opportunities facilitating promotions. This was identified as a critical factor in reducing turnover rates. Besides that, a study conducted in [28] aimed to identify the causes of employee turnover and its effects on the Saudi business environment. The study focused on factors such as job satisfaction, organizational commitment, support, feedback, rewards, and recognition. The results revealed that the primary factors influencing employees' decisions to leave included organizational support, rewards, appreciation, and commitment, whereas job satisfaction and

supervisor support had a comparatively less impact on turnover intention.

In contrast, the authors in [29] conducted a study to identify the employer retention methods and the factors affecting retention from the employee's perspective. The study was carried out in private businesses in three Saudi Arabian cities: Jeddah, Rabigh, and Yanbu. The findings indicated that increased rewards and privileges are key retention strategies for businesses aiming to maintain their staff long-term. Another study [30] demonstrated a negative relationship between employee experience and turnover intention, suggesting that as employees gain more experience, their likelihood of leaving decreases. Additionally, age plays a significant role in turnover motivation, with older workers exhibiting a lower propensity to leave compared to younger employees.

D. Overview of Predictive Models for Employee Attrition

Predictive ML involves training algorithms on historical data to generate predictions for new, unseen data. By leveraging various factors—such as job performance, length of service, and other relevant indicators— HR managers can evaluate an employee's intention to leave their position. Consequently, ML models have been employed extensively for the prediction of employee attrition. Inspired by previous studies, combinations of several of these models and techniques have been applied in this study to the newly created dataset, and the results have been analyzed in order to obtain robust predictions.

For instance, this study [13] suggested that people analytics be used to predict employee attrition, with a focus on data quality over quantity. The study employed three datasets: the IBM HR Analytics Employee Attrition dataset, a simulated HR dataset based on Kaggle data, and a survey-based dataset on the reasons for attrition, and applied a variety of algorithms, including deep learning methods, traditional ML models: DT, SVM, LR, RF, and ensemble techniques like XGBoost, VC, and stacking ANNs. The results indicated that the VC outperformed other models across all datasets. In the same way, the study in [31] evaluated other algorithms including KNN, Naive Bayes,RF, and LR on IBM HR Analytics Employee Attrition dataset, reporting the highest accuracy of 89% for Naive Bayes.

In addition, the study in [32] used various ML and ensemble learning methods to predict employee attrition, comparing four ML techniques and three ensemble methods, implementing grid search for hyperparameter tuning. The results reveal that the RF outperformed other models with an accuracy of 95%, identifying job satisfaction and stock options as primary attrition drivers. Similarly, Garg et al in [33] trained RF, AdaBoost, XGBoost, and ensemble stacking models on the IBM dataset, reporting the RF classifier as the top performing with 87.41% accuracy and 88% precision, and indicating that monthly pay significantly influences employee turnover.

In study [34], the IBM HR Analytics Employee Attrition and Performance dataset was used to train multiple models, including LR, RF, SVM, KNN, and DT. The LR model achieved the highest accuracy, with 87% precision and an F1 score of 83%. The findings identified monthly income, overtime, years with the current manager, age, distance from home, total working years, and years at the company as key attrition indicators. As well as previous research, [35] has used the IBM dataset available to evaluate LR, DT, RF, GBM, XGBoost, SVM, and KNN for attrition prediction. As a result, RF and GBM demonstrated superior predictive performance for employee retention. However, the SVM model achieved an accuracy of 88.78% and maintained a balanced performance with respect to accuracy and recall for the minority class (employees who departed), highlighting its usefulness in attrition prediction.

E. Research Gap and Study Positioning

Despite the growing body of research on employee attrition prediction using ML, several critical gaps remain that limit the practical applicability and contextual relevance of existing studies.

1) First: The majority of prior research has relied heavily on synthetic datasets, particularly the IBM HR Analytics Employee Attrition dataset [31], [33], [34], [35], or simulated data from platforms like Kaggle [13], [32]. While these datasets provide standardized benchmarks for model comparison, they fail to capture the complexity and nuances of real-world organizational dynamics, cultural contexts, and region-specific employment practices. This dependency on generic datasets raises questions about the generalizability and practical utility of the developed models in diverse organizational settings.

2) Second: There is a notable scarcity of research examining employee attrition within the Saudi Arabian context, despite the Kingdom's unique labor market characteristics shaped by Saudization policies, cultural factors, and rapid economic transformation [27], [28], [29]. The limited studies that do focus on Saudi Arabia primarily employ traditional statistical methods or exploratory analyses rather than comprehensive machine learning frameworks, leaving a gap in predictive analytics tailored to this specific workforce.

3) Third: Existing attrition prediction studies often utilize incomplete or inconsistent feature sets, omitting important variables identified in recent retention literature. Variables such as job opportunities, emotional commitment, health issues, job stability, and sector-specific factors are frequently overlooked in predictive models [27], [28], [31], despite evidence of their significance in employee turnover decisions.

4) Fourth: While ensemble methods have shown promise in attrition prediction [31], [32], [33], systematic comparisons across multiple feature selection techniques and their impact on model performance remain limited. Most studies employ a single feature selection approach without exploring how different methods influence predictive accuracy and model interpretability, particularly in the context of imbalanced datasets typical of attrition scenarios.

5) Finally: There is insufficient emphasis on deriving actionable insights for HR management from predictive models. Many studies focus primarily on achieving high prediction accuracy without adequately analyzing which factors most strongly influence attrition or how organizations can leverage these insights to develop targeted retention strategies [7], [14].

This study addresses these gaps by: 1) introducing a novel, empirically-collected dataset from real employees in the Saudi private sector that reflects authentic organizational contexts

and cultural dynamics; 2) incorporating a comprehensive set of attrition factors derived from both international literature and Saudi-specific studies, including previously underexplored variables, making this dataset valuable for addressing a range of factors not available in other attrition datasets, such as health issues, benefits, medical insurance, job opportunities, and others; 3) systematically comparing multiple machine learning algorithms and ensemble methods across various feature selection strategies to identify optimal prediction configurations; 4) providing detailed feature importance analysis to generate actionable insights for HR practitioners; and 5) demonstrating the practical applicability of predictive analytics in addressing attrition challenges within the evolving Saudi labor market. By filling these gaps, this research contributes both methodologically and contextually to the field of employee attrition prediction, offering a robust framework adaptable to other regional and organizational contexts.

III. Methodology

The methodology is structured to address three key objectives: 1) collecting authentic employee data that captures Saudi-specific factors, 2) preparing and balancing the dataset to mitigate prediction bias, and 3) systematically evaluating multiple ML models to identify the optimal prediction approach. Each phase builds upon the previous one to ensure robust and reliable attrition predictions. The data collection phase establishes the empirical foundation through a targeted survey of 1,191 Saudi private sector employees. The preprocessing phase addresses data quality through outlier removal, categorical encoding, and SMOTE-based class balancing. The modeling phase implements a comprehensive evaluation framework comparing eight algorithms across seven feature selection strategies, with performance assessed using accuracy and F1 score metrics optimized for imbalanced classification tasks. Fig. 1 shows the basic phases of the methodology followed in this study.

A. Data Collection

Employee data related to the attrition factors examined in this study were collected through an online questionnaire, specifically designed and implemented using the Google Forms platform. This survey specifically targeted real employees working within the Saudi private sector, thereby addressing the limitations of relying on synthetic datasets often used in prior research. In total, 1200 responses were gathered. The responses were thoroughly filtered to exclude individuals who did not belong to the Saudi private sector, resulting in a final sample of 1191 participants. Additionally, all respondents were employees from a wide range of organizations spread across the various regions of the Kingdom of Saudi Arabia, thereby ensuring diverse geographic representation within the dataset.

Moreover, the questionnaire was conducted anonymously to maintain participant confidentiality. With regard to demographics, male respondents constituted 43%, while female respondents made up 57% of the total sample. Additionally, 515 respondents indicated that they had left their jobs, providing direct insight into employee attrition.

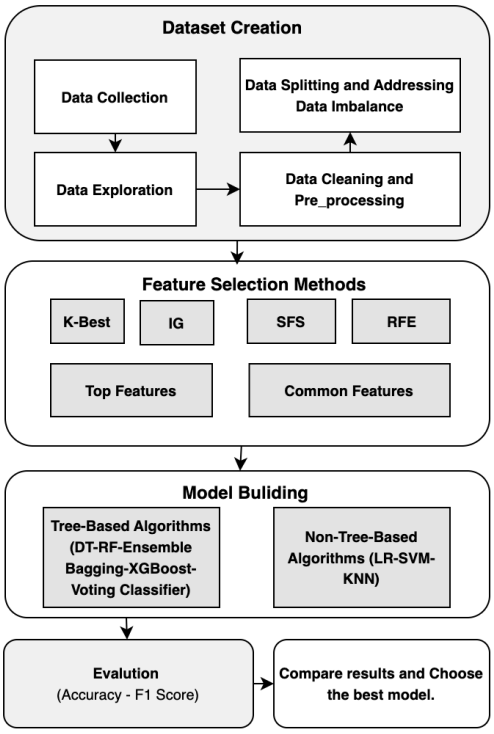


Fig. 1. Methodology phases.

TABLE I. Dataset Features

Category	Features
Personal Features	Gender, Age, Marital Status, Academic Degree, Health Issues, Years of Experience, Years of Experience in last organization, Job Opportunities.
Job-related Features	Sector, Department, Job Title, Monthly Salary, Allowances, Medical Insurance, Bonus, Overtime, Payment Overtime, Rewards and Wages Satisfaction, Get Deserved Promotion, Training Programs During the Last Three Years, Useful Training Programs, Business Travel, Job Support, Recognition, Emotional Commitment, Job Engagement, Distance to Work.
Psychological Features	Physical Stress, Psychological Exhaustion, Job Stability, Environment Satisfaction, Job Satisfaction , Work-Life Balance.
Target Feature	Attrition

B. Data Exploration

The dataset used in this study consists of 34 attributes, which are categorized into three primary groups: personality attributes, job-related attributes, and psychological and satisfaction attributes. First, personality attributes include individual characteristics such as gender, age, and social status. Second, job-related attributes encompass factors related to the work itself, such as job title, salary, sector, and department. Finally, psychological and satisfaction attributes relate to how employees perceive their work environment, including job stability, environmental satisfaction, and overall job satisfaction. The dataset also features the target variable of attrition, as detailed in Table I.

The variables included in this study were primarily selected based on insights from previous research on employee attrition prediction. In addition to these established factors,

the dataset integrates a set of novel variables that have not been widely applied in attrition prediction models. These novel variables originate from studies specifically focusing on the Saudi private sector, as well as recommended attrition predictors in the literature. For instance, key examples of these additional variables include job opportunities, allowances, job support, medical insurance, deserved promotion [27], [28], emotional commitment, health issues, and job stability [13]. Correspondingly, the sector variable was included to reflect the complex and diverse composition of the Saudi private sector, which encompasses a wide range of industries and organizational forms.

In this study, data exploration involved visualizing the distributions of key features, identifying correlations among variables, and detecting potential outliers or anomalies. To better understand the relationships between various factors and employee attrition, a thorough exploratory analysis was performed. The detailed distributions of the main features related to attrition are presented and further evaluated in the discussion section.

C. Data Cleaning and Pre-processing

The dataset contains both numerical and categorical features. Numerical variables included features such as employee allowances, whereas the remaining attributes were categorical in nature. In this section, the steps taken to preprocess the dataset are outlined to build ML models for employee attrition prediction.

1) Handling missing values: The dataset was examined for the presence of missing entries using the `isnull().sum()` function in Pandas. As no missing values were detected, neither imputation procedures nor the elimination of incomplete records was required.

2) Outlier detection and removal: To mitigate the influence of anomalous data points on model training, numerical features were evaluated for outliers using the Z-score technique. Instances with an absolute Z-score greater than three were classified as outliers and excluded from the dataset. This step ensured that extreme values did not distort the learning process of the predictive models.

3) Addressing categorical features categorical attributes processed using a two-stage approach:

a) Standardization: Preliminary exploration of the dataset revealed inconsistencies in the representation of categorical variables due to irregular spacing and capitalization. These issues were rectified by removing extraneous spaces and enforcing uniform formatting across feature values. This standardization process was applied to attributes such as `Business_Travel`, `Years_Experience`, and `MonthlySalary`.

b) Grouping rare categories: Several categorical variables, including `MonthlySalary`, `Sector`, `Years_experience_lastorganization`, `Years_Experience`, `Age`, and `Department` exhibited long-tailed distributions characterized by a large number of infrequent categories. To reduce model complexity and mitigate the risk of overfitting, these rare categories were consolidated into broader, more meaningful groups.

4) Encoding categorical variables: As a preliminary step, the LabelEncoder and OrdinalEncoder were employed to encode nominal and ordinal features, respectively. Due to the overfitting of tree-based models, in particular, the dataset was prepared in two distinct versions, with the key distinction being the encoding of categorical variables. In both dataset versions, ordinal features were encoded using Ordinal Encoding to preserve the intrinsic order among categories. Similarly, Label Encoding was consistently applied to binary features across both versions to ensure uniform representation.

For the tree-based dataset, nominal features with multiple categories, including `Maritalstatus`, `Sector`, `Department`, and `JobTitle` were encoded using the M-Estimator method. This encoder assigns numerical representations to categories, considering both the global distribution of the target variable and the distribution within each category [36]. By providing smooth and target-aware encodings, M-Estimator encoding enables tree-based models to create more effective and informative splits.

On the other hand, for the same nominal features, one-hot encoding was applied in the dataset version for non-tree-based models. This encoder generates binary dummy variables for each category, where the presence of a category is represented as 1 and its absence as 0, providing an orthogonal and non-ordinal representation suitable for these algorithms [37].

5) Scaling numerical features: Numerical and encoded features were normalized to ensure that the models could converge more quickly and avoid biasing the models towards features with larger ranges.

D. Data Splitting and Addressing Data Imbalance

The target variable, `Attrition`, exhibited significant class imbalance, with a notably higher proportion of instances labeled as "No" (employees who remained) compared to "Yes" (employees who left), as illustrated in Fig. 2. The dataset was partitioned into training and testing subsets, with an 80 %-20% split respectively, while maintaining the class distribution within each subset as detailed in Table II.

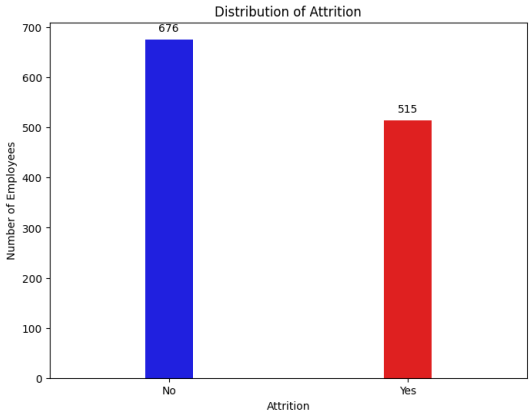


Fig. 2. Attrition distribution.

To address the class imbalance in the training dataset, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training set, targeting the minority

TABLE II. DATASET SPLIT AND ATTRITION DISTRIBUTION

#	Total Records	Non-Attrition	Attrition
Total records	1174	676	498
Training set 80%	939	534	405
SMOTE set	1068	534	534
Testing set 20%	235	142	93

class labeled "Yes". SMOTE generates synthetic samples by interpolating between existing minority class instances, thereby balancing the class distribution. This approach mitigates bias towards the majority class, enhances model performance on the minority class, and exposes the model to a more diverse and balanced set of examples, improving its generalization capability.

E. Feature Selection Methods

Feature selection was performed to identify the most informative features for predicting employee attrition in both tree-based and non-tree-based datasets. Four methods—SelectKBest, Mutual Information Gain (IG), Recursive Feature Elimination (RFE), and Sequential Feature Selection (SFS)—were applied, testing subsets of the top 10, 20, 25, and 30 features. The model performance was best with the top 25 features, consistent with prior studies in [38]. Therefore, this subset was selected for model training and evaluation.

F. Models Training and Validation

The modeling process aimed to predict employee attrition using a combination of tree-based and non-tree-based ML algorithms. This subsection details the models employed, evaluation metrics, and the overall procedures for training and testing

1) Feature sets: Seven feature sets were used to evaluate each model, reflecting various feature selection strategies. Each of these feature sets was evaluated to determine the optimal set of features for the models. These sets included: 1) all features without selection, 2) features from the SelectKBest method, 3) Information Gain-selected features, 4) Recursive Feature Elimination (RFE) features, 5) Sequential Feature Selection (SFS) features, 6) the union of features selected by all methods (common features), and 7) the intersection of features selected by any method (top features). This comprehensive evaluation allowed assessing model sensitivity to various feature subsets and selecting the most predictive features.

2) Model building: In this section, we describe the ML models used for predicting employee attrition, including both non-tree-based and tree-based algorithms. The non-tree-based algorithms employed in this study are listed in Table III, which includes LR, SVM, and KNN. These models are well-suited for datasets with one-hot encoded categorical features. Additionally, tree-based algorithms, as presented in Table IV, were utilized to leverage their ability to handle complex feature interactions and non-linear relationships. The tree-based models include DT, RF, Bagging ensemble, XGBoost, and a Voting Classifier combining Random Forest and XGBoost.

TABLE III. NON-TREE-BASED ALGORITHMS

No.	Model
1	Logistic Regression
2	Support Vector Machine (SVM)
3	K-Nearest Neighbors (KNN)

TABLE IV. TREE-BASED ALGORITHMS

No.	Model
1	Decision Tree
2	Random Forest
3	Bagging
4	XGBoost (Extreme Gradient Boosting)
5	Voting Classifier (Random Forest + XGBoost)

3) Evaluation metrics: Accuracy measures the proportion of correctly classified instances out of the total instances. It is calculated as:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}$$

Accuracy provides a simple and intuitive measure of the model's overall correctness. However, it can be misleading when dealing with imbalanced datasets.

The F1 Score is defined as the harmonic mean of precision and recall, providing a quantitative measure that balances these two metrics. This metric is formally expressed as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score is especially advantageous when it is necessary to account for both false positives and false negatives. A higher F1 score indicates robust model performance across both dimensions. Notably, the F1 score often provides a more informative evaluation than accuracy in scenarios characterized by class imbalance, due to reflecting the model's proficiency with respect to the minority.

IV. RESULTS

A. AL Models with Different Features Selection Methods for the Top 25 Features

The performance of various ML models using different feature selection strategies was evaluated. The models evaluated include LR, DT, RF, XGBoost, Ensemble Bagging, SVM, KNN, and VC. Each model was assessed using seven feature selection strategies: All Features, SelectKBest, IG, RFE, SFS, Common Feature, and Top Feature. These strategies utilized the top 25 ranked features based on their respective scores. After training under all feature selection scenarios, results were compared to identify the optimal configuration per model. For each model, only the best result and corresponding feature set are reported as shown in Table V, which summarizes the best performance of each model with feature sets. Additionally, the highest accuracy observed in this study was 90%, and the highest F1 score was 85%, both achieved by the Voting Classifier leveraging the feature set selected by RFE.

TABLE V. MODELS PERFORMANCE

Model	Fatures selection	Accuracy	F1 Score
LR	All	77.00	74.00
SVM	RFE	80.34	71.00
K-NN	Common	78.63	75.25
DT	Common	84.62	76.32
RF	SFS	89.00	84.71
Ensemble Bagging	SFS	87.18	82.35
XGBoost	All	83.76	75.95
Voting Classifier	RFE	**90.00**	**85.00**

B. Comparative Analysis with Existing Methods

To validate the effectiveness of the proposed approach and assess its importance relative to existing methods, we compare our best-performing model (Voting Classifier with RFE features) against results reported in recent studies. Table VI presents this comparison based on the methodologies and results explicitly reported in the referenced studies.

The comparative analysis demonstrates that the proposed method achieves competitive performance, with the Voting Classifier achieving 90% accuracy and 85% F1 score. This performance is comparable to the best results reported in existing literature, with several important distinctions.

First, unlike prior studies that predominantly utilize the IBM HR Analytics Employee Attrition dataset [31], [33], [34], [35], our approach employs a novel, empirically-collected dataset from the Saudi private sector, which introduces additional complexity due to real-world data characteristics and regional-specific factors. Second, our method achieves this performance while addressing class imbalance through SMOTE and systematic feature selection across multiple methods (SelectKBest, IG, RFE, SFS), providing a more comprehensive evaluation framework than many existing approaches. Third, the high F1 score (85%) indicates strong performance on both precision and recall, which is particularly important for imbalanced attrition datasets where the minority class (employees who leave) is of primary interest.

The study by Yahia et al. [13] also reported that Voting Classifier outperformed other models across multiple datasets, which aligns with our findings and validates the effectiveness of ensemble methods for attrition prediction. However, our study extends this by demonstrating the applicability of these methods to region-specific, real-world data and by providing detailed feature importance analysis that reveals job opportunities and job titles as primary attrition drivers in the Saudi context.

The effectiveness of hybrid ensemble approaches is further demonstrated by Govindarajan et al. [39], whose Hybrid Model integrating multiple algorithms achieved 95% accuracy on the IBM HR dataset. Their comparative analysis revealed that individual models achieved lower performance (SVM: 88.6%, Random Forest and Gradient Boosting: 87.3%, Decision Tree: 80.5%), while the hybrid approach effectively harnessed complementary strengths while mitigating individual weaknesses. This finding directly supports our methodology of employing a Voting Classifier that combines Random Forest and XGBoost, demonstrating that ensemble integration consistently outperforms individual algorithms across different studies and datasets. The parallel between their Hybrid Model (95%) and our Voting Classifier (90%) performance, despite our use of more challenging real-world data, validates the robustness of ensemble-based approaches for attrition prediction.

V. DISCUSSION

This study addresses the pressing issue of employee attrition within the Saudi private sector, with a focus on developing predictive models using ML techniques. The research aims to determine how well ML models can predict employee attrition, and to identify the feature configurations that most effectively enhance prediction accuracy. The subsequent discussion synthesizes the key findings in relation to the research questions, objectives, and methodology introduced earlier.

A. Discussion Related to Dataset

This study utilized a new, real-world dataset that stands out for its relevance, depth, and contextual specificity to the Saudi Arabian private sector workforce. Unlike many prior works relying on synthetic or general-purpose datasets such as IBM's attrition data [31], [33], [34], and [35], this research provides a more realistic and localized perspective. The dataset comprises responses from 1,191 participants gathered through an online questionnaire targeting Saudi private-sector employees across various industries. The dataset's contextual specificity enables the extraction of actionable insights tailored to the distinct work culture and employment practices prevalent in Saudi Arabia, which often diverge from Western contexts. A significant insight from this dataset is the class imbalance in the target variable: approximately 43% of participants had left a job, while 57% had not. To address this imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was applied during training, which is critical for mitigating prediction bias in high-stakes human resource analytics. However, the geographic and cultural specificity of the dataset may constrain the generalizability of the findings to other labor markets.

B. Discussion Related to Employee Attrition Factors

A variety of feature selection techniques were employed in this study. Due to the multiplicity and potential conflict among selected feature groups, a general, model-agnostic approach—binary logistic regression—was utilized to assess the predictive significance of variables. Logistic regression offers interpretable correlation measures in the form of odds ratios, which provide straightforward insights compared to the often complex interpretability of ML outputs [40]. In addition, a subset of common variables was identified as the intersection of all applied feature sets, representing features consistently deemed significant across methods.

The results of the logistic regression analysis revealed that the most significant factors influencing employee attrition were job opportunities, job title, deserved promotion, training programs, emotional commitment, job engagement, benefits, and medical insurance. The common set of attributes selected from all attribute selection methods included the following variables: years of experience, job opportunities, sector, job engagement, job support, years of experience in the previous organization, job stability, training programs, appreciation, emotional commitment, health issues, and deserved promotion.

TABLE VI. COMPARISON WITH EXISTING EMPLOYEE ATTRITION PREDICTION METHODS

Study	Dataset	Best Model	Accuracy	F1 Score
Rana et al. [31]	IBM HR Dataset	Naive Bayes	89%	–
Chung et al. [32]	IBM HR Dataset	Random Forest	95%	–
Garg et al. [33]	IBM Dataset	Random Forest	87.41%	88% (Precision)
Solomon et al. [34]	IBM HR Dataset	Logistic Regression	87%	83%
Amzad et al. [35]	IBM Dataset	SVM	88.78%	–
Yahia et al. [13]	Large, medium simulated HR datasets, small (450 real employees)	Deep and ensemble learning	96%, 98%, 99% respectively	–
Proposed Method	**Saudi Private Sector (1,191 real employees)**	**Voting Classifier (RFE)**	**90%**	**85%**

The results align with prior studies investigating employee attrition determinants in the Saudi private sector. Study [27] highlighted the significant influence of job opportunities, training programs, and promotion on attrition rates. In the same way, the study in [13] underscored the impact of job title and job engagement, emphasizing additional factors such as job stability, health conditions, and emotional commitment. Moreover, study [28] affirms the role of job support and recognition in reducing attrition. Furthermore, study [29] identified job support, recognition, financial allowances, and bonuses as critical factors in employee retention consistent with. The importance of years of experience in predicting employee attrition was also highlighted in [34]. On the other hand, several variables were not statistically significant in predicting workforce attrition, including age, academic degree, monthly salary, overtime hours, bonuses, job satisfaction, environmental satisfaction, business travel, and distance to work, which contradicts previous studies [13], [34] . These contradictions reflect contextual differences in the Saudi private sector. The non-significance of salary and bonuses suggests competitive compensation is a baseline expectation rather than a differentiator, shifting focus to career growth. Age's reduced importance aligns with Saudi Arabia's young workforce in Vision 2030 transformation [30], where career prospects uniformly influence turnover. Job satisfaction's non-significance indicates employees may be satisfied yet leave for better opportunities, suggesting satisfaction alone is insufficient without development prospects. The non-significance of business travel contradicts Yahia et al. [13], likely reflecting cultural and work-life balance differences in the Saudi context. These findings emphasize that retention strategies from Western datasets require adaptation to Middle Eastern labor markets.

C. Discussion Related to ML Models

1) ML Models: Based on the study results, the voting classifier with the feature set selected by Recursive Feature Elimination (RFE) achieved the best performance, with 90% accuracy and an 85% F1 score. This finding aligns with the results reported in study [13]. Ensemble learning methods, such as voting classifiers, combine multiple weak learners to leverage their complementary strengths, resulting in more accurate predictions. Combining outputs from multiple models is a well-established technique to improve overall model performance [41]. Apart from this, RF achieved the second-best performance, consistent with its superiority in previous studies [32],[33], and [35]. Conversely, the LR model did not outperform the other models, contrasting with findings reported in [38].

2) Impact of Feature Sets: The "All Features" set includes all available features for comparison, showing good overall performance. Specifically, the LR achieved the best result with all features with 77% accuracy and 74% F1 score. The evaluation of "Common Features" across models revealed varying impact: while the accuracy of VC and RF improved, the performance of LR, SVM, and Bagging declined significantly. Besides that, the "Top Features" method consistently improved model performance, although it was less effective compared to SFS and RFE.

In contrast, most models demonstrated improved performance with feature selection methods. Models such as SVM, RF, and VC benefited from feature selection by reducing noise and focusing on only relevant variables. SFS emerged as the most effective feature selection method for most models. For instance, with ensemble bagging, SFS provided the highest accuracy at 87.18%. SFS emerged as the most effective technique overall, providing the highest accuracy of 87.18% for the ensemble bagging model. Additionally, RFE ranked as the second-best feature selection method after SFS, improving model performance, particularly for RF, Bagging, and VC models.

While SelectKBest and IG effectively identified important features, their impact on improving model performance was less significant compared to SFS and RFE. For example, Logistic Regression (LR) achieved an F1 score of 73.58% using KBest, which increased to 79.31% with SFS. Similarly, Random Forest (RF) obtained an F1 score of 73.58% with IG, which improved to 84.71% using SFS and reached 80% with RFE as shown in detail in Fig. 3, which presents a comparison of accuracy across various models and feature selection methods.

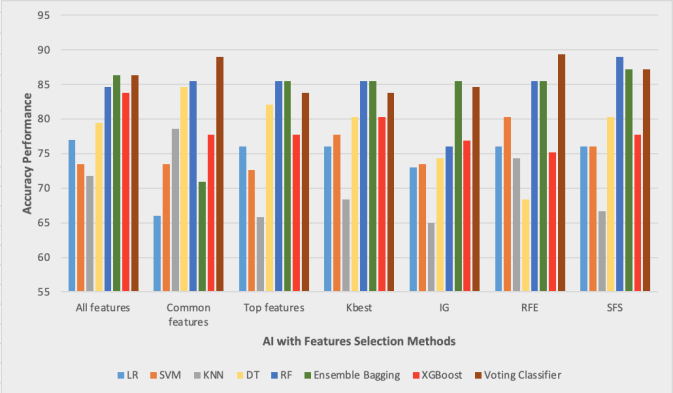


Fig. 3. Comparison of AI model accuracy performance across different feature sets and feature selection methods.

VI. Conclusion

Employee attrition remains a critical challenge for the private sector, particularly in monitoring employee intentions and proactively addressing turnover. This study aimed to analyze key factors predicting employee departure and the underlying reasons for turnover, thereby enabling organizations to implement effective retention strategies. The primary objective of this study was to develop and evaluate a predictive model using ML and ensemble methods to address attrition challenges tailored to the Saudi private sector workforce.

The study has four key contributions. First, it introduces a new dataset derived from previously identified causes of employee attrition. Next, it presents the design, development, and evaluation of predictive models using ML and ensemble methods. Additionally, the study compares the performance of various ML models across different feature configurations. Lastly, it provides a thorough analysis of factors affecting turnover intentions among employees in the Saudi private sector, contributing valuable insights to inform retention policy development.

The study results indicate the superiority of ensemble models over individual algorithms in predicting employee attrition. Furthermore, demographic and overall satisfaction variables showed limited predictive power within this dataset, while dynamic and developmental factors had a more pronounced influence on attrition. These findings reveal an important shift in employee attrition patterns: traditional demographic-based retention strategies may be insufficient in contemporary organizational contexts. Instead, the prominence of dynamic factors—such as job opportunities, career development prospects, and job titles—suggests that employees in the Saudi private sector prioritize growth-oriented and opportunity-driven workplace environments. This insight challenges conventional HR assumptions that focus primarily on compensation and job satisfaction, highlighting the need for organizations to invest in career development programs, clear advancement pathways, and competitive positioning within the job market. in additon to, need to adopt proactive strategies focused on development and career advancement. The proposed model can help identify at-risk employees early and support data-driven retention decisions in the Saudi private sector.

From a strategic perspective, the high predictive accuracy of ensemble models (90% with Voting Classifier) demonstrates the viability of data-driven early warning systems for attrition risk. Organizations can leverage such models to identify at-risk employees proactively, enabling targeted interventions before turnover intentions crystallize into actual departures. This predictive capability is particularly valuable for retaining high-performing employees and reducing the substantial costs associated with talent replacement and organizational knowledge loss.

However, the study faced limitations, including potential bias inherent to questionnaire-based data collection, while relying on actual employee records in private sector companies may provide more reliable data. Additionally, the focus on the Saudi private sector as a whole suggests that sector-specific studies—for example, healthcare or education—could yield more precise insights into attrition drivers.

Future research should pursue several promising directions to advance employee attrition prediction and understanding. First, addressing the challenge of imbalanced data typical of turnover datasets through advanced techniques such as SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic Sampling), or cost-sensitive learning algorithms could improve model performance and generalizability. Second, exploring deep learning architectures, including recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, may capture temporal patterns in employee behavior and career trajectories, potentially revealing early warning signals that precede turnover decisions. Third, incorporating explainable AI (XAI) techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) would enhance model interpretability, enabling HR practitioners to understand not only predictions but also the reasoning behind them, thereby facilitating more informed decision-making. Fourth, conducting longitudinal studies that track employees over extended periods could reveal the dynamic evolution of attrition risk factors and validate the temporal stability of predictive models. Finally, sector-specific investigations within healthcare, education, technology, and finance sectors would uncover industry-unique attrition drivers and enable the development of tailored retention strategies.

Ultimately, this study offers a foundational step toward enhancing employee retention strategies within Saudi Arabia's evolving private sector landscape, demonstrating that the integration of machine learning techniques with empirical workforce data can transform HR practices from reactive problem-solving to proactive talent management.

References

[1] K. M. Mitravinda and S. Shetty, "Employee attrition: Prediction, analysis of contributory factors and recommendations for employee retention," in *2022 IEEE Int. Conf. Women Innovation, Technol. Entrepreneurship (ICWITE)*, Dec. 2022, pp. 1–6.

[2] F. Mozaffari, M. Rahimi, H. Yazdani, and B. Sohrabi, "Employee attrition prediction in a pharmaceutical company using both machine learning approach and qualitative data," *Benchmarking: An Int. J.*, 2022, ahead-of-print.

[3] R. Joseph, S. Udupa, S. Jangale, K. Kotkar, and P. Pawar, "Employee attrition using machine learning and depression analysis," in *2021 5th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*. IEEE, 2021, pp. 1000–1005.

[4] S. Gupta, G. Bhardwaj, M. Arora, R. Rani, P. Bansal, and R. Kumar, "Employee attrition prediction in industries using machine learning algorithms," in *2023 10th Int. Conf. Comput. Sustainable Global Develop. (INDIACom)*. IEEE, 2023, pp. 945–950.

[5] G. R. Rajeswari, R. Murugesan, R. Aruna, B. Jayakrishnan, and K. Nilavathy, "Predicting employee attrition through machine learning," in *2022 3rd Int. Conf. Smart Electron. Commun. (ICOSEC)*. IEEE, 2022, pp. 1022–1027.

[6] E. Rombaut and M.-A. Guerry, "Predicting voluntary turnover through human resources database analysis," *Manage. Res. Rev.*, vol. 41, no. 1, pp. 96–112, 2018.

[7] N. B. Yahia, J. Hlel, and R. Colomo-Palacios, "From big data to deep data to support people analytics for employee attrition prediction," *IEEE Access*, vol. 9, pp. 60 447–60 458, 2021.

[8] A. I. Al-Alawi and Y. A. Ghanem, "Predicting employee attrition using machine learning: A systematic literature review," in *2024 ASU Int. Conf. Emerging Technol. Sustainability Intell. Syst. (ICETSIS)*. IEEE, Jan. 2024.

[9] A. Tursunbayeva, S. D. Lauro, and C. Pagliari, "People analytics—A scoping review of conceptual boundaries and value propositions," *Int. J. Inf. Manage.*, vol. 43, pp. 224–247, 2018.

[10] D. Angrave, A. Charlwood, I. Kirkpatrick, M. Lawrence, and M. Stuart, "HR and analytics: Why HR is set to fail the big data challenge," *Human Resource Manage. J.*, vol. 26, no. 1, pp. 1–11, 2016.

[11] T. Pape, "Prioritising data items for business analytics: Framework and application to human resources," *Eur. J. Oper. Res.*, vol. 252, no. 2, pp. 687–698, 2016.

[12] P. K. Jain, M. Jain, and R. Pamula, "Explaining and predicting employees' attrition: A machine learning approach," *SN Appl. Sci.*, vol. 2, pp. 1–11, 2020.

[13] N. B. Yahia, J. Hlel, and R. Colomo-Palacios, "From big data to deep data to support people analytics for employee attrition prediction," *IEEE Access*, vol. 9, pp. 60 447–60 458, 2021.

[14] T. Peeters, J. Paauwe, and K. V. D. Voorde, "People analytics effectiveness: Developing a framework," *J. Organizational Effectiveness: People Performance*, vol. 7, no. 2, pp. 203–219, 2020.

[15] H. Alqahtani, H. Almagrabi, and A. Alharbi, "Employee attrition prediction using machine learning models: A review paper," *Int. J. Artif. Intell. Appl.*, vol. 15, no. 2, 2024.

[16] A. Raza, K. Munir, M. Almutairi, F. Younas, and M. M. S. Fareed, "Predicting employee attrition using machine learning approaches," *Appl. Sci.*, vol. 12, no. 13, p. 6424, 2022.

[17] A. Mhatre, A. Mahalingam, M. Narayanan, A. Nair, and S. Jaju, "Predicting employee attrition along with identifying high risk employees using big data and machine learning," in *2020 2nd Int. Conf. Adv. Comput., Commun. Control Netw. (ICACCCN)*. IEEE, 2020, pp. 269–276.

[18] S. Dutta, S. K. Bandyopadhyay, and S. K. Bandyopadhyay, "Employee attrition prediction using neural network cross validation method," *Int. J. Commerce Manage. Res.*, vol. 6, no. 3, pp. 80–85, 2020.

[19] K. Ashokkumar, S. Jacob, and A. Joseph, "A study on attrition management in private sector financial institutions—A special reference to Kottayam District in Kerala," 2020, technical report.

[20] D. S. Sisodia, S. Vishwakarma, and A. Pujahari, "Evaluation of machine learning models for employee churn prediction," in *2017 Int. Conf. Inventive Comput. Inform. (ICICI)*. IEEE, 2017, pp. 1016–1020.

[21] D. Alao and A. B. Adeyemo, "Analyzing employee attrition using decision tree algorithms," *Comput., Inf. Syst., Develop. Inform. Allied Res. J.*, vol. 4, no. 1, pp. 17–28, 2013.

[22] F. H. Wardhani and K. M. Lhaksmana, "Predicting employee attrition using logistic regression with feature selection," *Sinkron: Jurnal dan Penelitian Teknik Inform.*, vol. 7, no. 4, pp. 2214–2222, 2022.

[23] J. T. Johnson, R. W. Griffeth, and M. Griffin, "Factors discriminating functional and dysfunctional salesforce turnover," *J. Bus. Ind. Marketing*, 2000.

[24] J. Scott, S. Waite, and D. Reede, "Voluntary employee turnover: A literature review and evidence-based, user-centered strategies to improve retention," *J. Amer. College Radiol.*, vol. 18, no. 3, pp. 442–450, 2021.

[25] R. Basariya and R. R. Ahmed, "A study on attrition—turnover intentions of employees," *Int. J. Civil Eng. Technol.*, vol. 10, no. 1, pp. 2594–2601, 2019.

[26] P. Ghosh, R. Satyawadi, J. P. Joshi, and M. Shadman, "Who stays with you? Factors predicting employees' intention to stay," *Int. J. Organizational Anal.*, vol. 21, no. 3, pp. 288–312, 2013.

[27] S. Muzaffar and U. Javed, "Training and development opportunities and turnover intentions post saudization," *PalArch's J. Archaeology Egypt/Egyptology*, vol. 18, no. 14, pp. 521–531, 2021.

[28] W. Bashraheel and M. Abunar, "The factors of employee turnover intention in private sectors organization in Jeddah, Saudi Arabia," *PalArch's J. Archaeology Egypt/Egyptology*, vol. 18, no. 12, pp. 87–94, 2021.

[29] R. H. Al-Gharbi and U. Javed, "Factors impacting employee retention: A case of private companies in Saudi Arabia," *PalArch's J. Archaeology Egypt/Egyptology*, vol. 18, no. 14, pp. 771–778, 2021.

[30] S. Faisal, M. Naushad, and M. Faridi, "A study on the level and relationship of job embeddedness and turnover intentions among Saudi Arabian working-class," *Manage. Sci. Lett.*, vol. 10, no. 13, pp. 3167–3172, 2020.

[31] A. Rana, S. Malik, and M. Chauhan, "Employee attrition prediction using ML techniques," *J. Kufa Math. Comput.*, vol. 11, no. 2, pp. 88–100, Aug. 2024.

[32] D. Chung, J. Yun, J. Lee, and Y. Jeon, "Predictive model of employee attrition based on stacking ensemble learning," *Expert Syst. Appl.*, vol. 215, p. 119364, 2023.

[33] U. Garg, N. Gupta, M. Manchanda, and K. Purohit, "Classification and prediction of employee attrition rate using ML classifiers," in *2024 Int. Conf. Inventive Comput. Technol. (ICICT)*. IEEE, Apr. 2024, pp. 608–613.

[34] C. H. Solomon, D. Mohankumar, and C. Sivanandam, "Employee attrition analysis using ML," in *2024 Second Int. Conf. Emerging Trends Inf. Technol. Eng. (ICETITE)*. IEEE, Feb. 2024, pp. 1–8.

[35] M. S. A. Basha, O. Varikunta, A. U. Devi, and S. Raja, "ML in HR analytics: A comparative study on the predictive accuracy of attrition models," in *2024 2nd Int. Conf. Device Intell., Comput. Commun. Technol. (DICCT)*. IEEE, Mar. 2024, pp. 475–480.

[36] D. Micci-Barreca, "A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems," *SIGKDD Explor.*, vol. 3, pp. 27–32, Jul. 2001.

[37] Various Authors, "Encoding categorical data: Is there yet anything 'hotter' than one-hot encoding?" *arXiv preprint arXiv:2312.16930*, 2023, available: https://ar5iv.org/abs/2312.16930.

[38] S. F. Sari and K. M. Lhaksmana, "Employee attrition prediction using feature selection with information gain and random forest classification," *J. Comput. Syst. Inform. (JoSYC)*, vol. 3, no. 4, pp. 410–419, 2022.

[39] R. Govindarajan, N. K. Kumar, P. S. Reddy, E. S. Pravallika, B. Dhatri, and G. P. Kumar, "Predicting employee attrition: A comparative analysis of machine learning models using the IBM human resource analytics dataset," *Procedia Comput. Sci.*, vol. [volume], p. [pages], 2025.

[40] H. S. R. Rajula, G. Verlato, M. Manchia, N. Antonucci, and V. Fanos, "Comparison of conventional statistical methods with ML in medicine: Diagnosis, drug development, and treatment," *Medicina*, vol. 56, no. 9, p. 455, Sep. 2020, pMID: 32911665; PMCID: PMC7560135.

[41] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Eng. Appl. Artif. Intell.*, vol. 115, p. 105151, 2022.

Optimizing Asset Transfer Process in ERP Using Business Process Management Technique

Ravindu Yasarathne[1], Naduni Ranatunga[2], Vikasitha Herath[3],
Lakshan Chalinda[4], Chathurangika Kahandawaarachchi[5], Sanjeeva Perera[6], Chamath Randula[7]
Department of Computer Systems Engineering-Faculty of Computing,
Sri Lanka Institute of Information Technology, Colombo, Sri Lanka[1,2,3,4,5,6]
SLIIT Business School, Sri Lanka Institute of Information Technology, Colombo, Sri Lanka[7]

***Abstract*—Enterprise Resource Planning (ERP) systems are critical for managing enterprise-wide business processes, including asset management. Yet, many ERP platforms lack efficient mechanisms for bulk asset transfers, leading to high manual effort, increased costs, and data inconsistencies. This study applies Business Process Reengineering (BPR) techniques as the methodology to optimize ERP asset management, focusing on workflow optimization and automation, contributing both practical and methodological insights. A mixed-method approach was adopted, analyzing a financial organization with 256 branches and over 450 Oracle ERP users. Data from 51 representative branches identified inefficiencies such as manual transfer delays, approval bottlenecks, and synchronization issues. The proposed solution introduces automated bulk asset transfers, optimized approval workflows, and real-time data synchronization, along with new metrics for evaluating efficiency, compliance, risk, and asset utilization. Compared to the As-Is system, the reengineered framework achieved a 100% reduction in operational costs per user ($7,500 annual saving), an 80% reduction in compliance incidents, a 67% reduction in asset transaction errors, and a 20% improvement in asset utilization. These results demonstrate a scalable, adaptable, and effective framework that enhances ERP operational efficiency, strengthens data integrity, and advances both academic understanding and industrial practice of asset management process reengineering.**

Keywords*—Asset management; bulk asset transfer; Business Process Reengineering (BPR); Enterprise Resource Planning (ERP); workflow optimization

TABLE I. ABBREVIATIONS

Abbreviations	Meaning
BPM	Business Process Management
BPR	Business Process Reengineering
ERP	Enterprise Resource Planning
GRN	Good Receive Note
IT	Information Technology
BPO	Business Process Outsourcing
KPI	Key Performance Indicator
UAT	User Acceptance Testing
PL/SQL	Procedural Language/Structured Query Language
BPMN	Business Process Model and Notation
AI	Artificial Intelligence
TQM	Total Quality Management
DMAIC	Define, Measure, Analyze, Improve, and Control
APEX	Oracle Application Express
HTML5	Hypertext Markup Language version 5
BPMM	Business Process Maturity Modeling
Forward Integration	Syncing data from Oracle fusion to the developed system
Reverse Integration	Real-time synchronization of data from developed system to Oracle fusion back

ABBREVIATIONS

Table I shows the abbreviations that are used in this manuscript.

I. INTRODUCTION

The current As-Is asset management process in Enterprise Resource Planning (ERP) systems faces several inefficiencies that hinder operational effectiveness. These inefficiencies arise primarily from the absence of a bulk asset transfer feature, which requires users to process asset transfers individually—making the process time-consuming and prone to human error. Manual interventions and repetitive tasks contribute to operational delays, increased labor costs, and data inconsistencies, necessitating a comprehensive overhaul through Business Process Reengineering (BPR).

Several studies and implementation reports have highlighted enhancements to the asset transferring process in Oracle Cloud ERP, aiming to improve efficiency and accuracy through automation and streamlined workflows. However, these contributions primarily focus on incremental improvements to existing processes rather than a holistic redesign. Moreover, the literature provides limited evidence regarding the effectiveness of these enhancements, with insufficient methodological details or empirical validation to demonstrate measurable performance gains. Consequently, there remains a significant research and practical gap in optimizing the asset transfer process itself—particularly in enabling bulk transfer capabilities. Addressing this gap through a reengineered, BPR-driven approach introduces a clear element of novelty by systematically enhancing efficiency, automation, and integration within the asset management lifecycle [1].

This strategic approach focuses on radically redesigning an organization's core processes to achieve significant improvements in performance, efficiency, and operational outcomes. In the context of asset management, BPR plays a crucial role in workflow optimization, reducing manual intervention, and enhancing data accuracy. By reengineering asset transfer processes, particularly in large organizations utilizing ERP systems, it becomes possible to efficiently manage complex asset movements across multiple locations.

A. Research Gap

Although ERP systems are designed to integrate and automate enterprise-wide processes, most lack built-in capabilities for bulk asset transfers and customizable approval workflows. Existing studies primarily focus on process automation or ERP customization in general, with limited attention to optimizing asset transfer workflows using BPR principles. This research fills that gap by developing and evaluating a reengineered, workflow-optimized bulk asset transfer mechanism integrated with real-time synchronization and enhanced approval processes within Oracle ERP.

B. Statement of the Research Problem

The current ERP asset management modules do not provide efficient mechanisms for bulk asset transfers, resulting in manual, time-intensive processes that cause operational inefficiencies, increased costs, and potential data inconsistencies. The lack of workflow automation and integration between asset movement and approval mechanisms further limits performance. Therefore, there is a need to reengineer the existing asset transfer process to enable automation, workflow optimization, and seamless integration within ERP environments.

C. Research Questions

- How can Business Process Reengineering (BPR) be applied to optimize the asset transfer process within ERP systems?
- What improvements in workflow efficiency, data accuracy, and operational cost can be achieved through the implementation of an automated bulk asset transfer feature?
- How can forward and reverse integration mechanisms enhance synchronization and data consistency in reengineered ERP modules?

D. Research Significance

This research contributes to both theory and practice by bridging a critical gap in ERP asset management. The proposed BPR-driven bulk asset transfer solution introduces a scalable framework for workflow optimization, automation, and integration across ERP systems. It benefits organizations by improving process efficiency, reducing manual workload, enhancing data accuracy, and providing a model that can be adapted by other enterprises facing similar limitations. The outcomes also support ERP vendors and practitioners in extending native ERP capabilities without additional licensing costs, promoting cost-effective digital transformation within enterprise environments.

E. Research Objectives

The primary objective of this research is to identify and reengineer inefficiencies in the existing asset management process within ERP systems in a selected financial institution.

The sub-objectives of this research include:

- To perform a detailed analysis of the As-Is process of asset transfer in the selected financial organization using the Oracle ERP system.
- To develop the proposed system using Oracle APEX and other technologies with forward and reverse integration mechanisms.
- To use Business Process Reengineering tools, we are going to perform a detailed analysis of the developed system for asset transfers in the selected financial organization using the Oracle ERP system.
- To identify the metrics for measuring the success of the developed system.

This study focuses on the reengineering of bulk asset transfers, reverse integration, and workflow automation within ERP systems. Additionally, it examines enhancements in user security and privilege management to ensure that access to sensitive asset-related information is restricted to authorized personnel. However, the scope of this research is confined to implementations within the ERP environment.

II. Literature Review

A. Business Process Reengineering (BPR) in ERP Systems

Enterprise software such as Enterprise Resource Planning (ERP) systems are essential for modern businesses as they integrate core functions such as finance, supply chain, HR, and customer management into a centralized platform. By automating workflows and improving data accuracy, ERP enhances operational efficiency, reduces costs, and supports better decision-making through real-time analytics [2]. ERP systems also provide scalability and adaptability, enabling enterprises to respond to market changes effectively. Their ability to streamline processes and ensure data consistency makes them critical tools for business growth and sustainability [3], [4].

Some studies explore cloud ERP using an integrative model combining the Technological, Organizational, and Environmental (TOE) framework and the Unified Theory of Acceptance and Use of Technology (UTAUT), which identifies key adoption factors such as system reliability, data security, cost efficiency, scalability, and accessibility. These insights are useful for optimizing ERP's asset transfer module, emphasizing security, efficiency, and seamless cloud integration [5].

Business Process Management (BPM) is an approach for improving and optimizing an organization's processes to achieve more efficient workflow and better outcomes. BPM involves analyzing, modeling, implementing, monitoring, and continuously improving business processes. By aligning processes with business objectives, the BPM enhances productivity, reduces costs, and improves agility. It often integrates various tools and methodologies, including automation, analytics, and performance metrics, to drive process improvements across departments and organizational levels. Effective Business Process Management (BPM) ensures a streamlined approach to achieving business goals and adapting to changes [6] Business Process Reengineering (BPR) is a branch of Business Process Management. It was designed to achieve significant improvements in efficiency and performance by fundamentally rethinking and redesigning core business processes. Rather than making incremental changes, BPR focuses on eliminating inefficiencies, optimizing workflows, and enhancing the overall operational effectiveness. This transformation has lead to

improvements in cost reduction, productivity, service quality, process speed, and overall business agility. Since business environments are constantly evolving, BPR is not a one-time initiative but an ongoing effort that requires continuous evaluation and adaptation to maintain competitive advantage [7].

The implementation of BPR begins with a comprehensive analysis of the existing processes to identify inefficiencies, bottlenecks, and areas for improvement. This analysis informs strategic decisions, such as process automation, workflow optimization, business process outsourcing (BPO), or redefining third-party roles. A key aspect of BPR is leveraging information technology (IT) to automate repetitive tasks, restructure workflows, and eliminate non-value-added activities, thereby enabling seamless process execution and improved decision-making [7].

By rethinking and redesigning business processes, organizations can modernize operations, enhance customer service, reduce operational costs, and improve their adaptability to market shifts. When implemented effectively, BPR enables enterprises to remain competitive in dynamic business environments by continuously optimizing their operational frameworks [8]. Evaluating the success of Business Process Reengineering (BPR) initiatives in Enterprise Resource Planning (ERP) systems requires a set of measurable key performance indicators (KPIs). The most commonly used metrics are "cost savings, process cycle time reduction, data accuracy, operational efficiency, and customer satisfaction". Among these metrics, the Business Process Reengineering (BPR) is mainly focuing on "cost reduction" through process efficiency. Organizations measure cost savings by analyzing reductions in labor costs, maintenance expenses, and inventory carrying costs [9]. The other main metric is the "Process Cycle Time Reduction" where it assesses the impact of Business Process Reengineering (BPR) on the speed of business operations. Companies experience faster processing times by automating workflows and eliminating redundant steps. As mentioned above, accurate real-time data are essential for decision-making. BPR initiatives often implement data validation rules, automation, and system integrations to minimize human errors. The effectiveness of these measures is evaluated by tracking data discrepancies, error rates in financial reporting, and compliance adherence within ERP systems [10]. Rather than the above metrics, organizations assess improvements in workforce productivity by analyzing reductions in manual effort and increases in task automation. Productivity gains are often measured using employee performance benchmarks, transaction processing speeds, and system response times after BPR implementation. Based on the assessment of the above KPIs, a successful BPR initiative should enhance customer experience by improving service delivery. Metrics such as response times, and customer feedback scores provide insight into how process reengineering influences end-user satisfaction [11].

The literature on ERP systems, asset management, and Business Process Reengineering (BPR) provides valuable insights into improving enterprise operations through automation and workflow optimization as mentioned above. Previous studies have focused on ERP implementation challenges, process automation, and performance enhancements, but they often address these topics in general terms without targeting the specific inefficiencies in asset transfer processes. Notably, most ERP platforms lack native support for bulk asset transfers, forcing organizations to rely on manual, time-consuming processes prone to errors and delays. While some research has explored workflow redesign or integration strategies, there is limited empirical evidence demonstrating the combined effect of workflow optimization, automation, and real-time data synchronization in reengineering ERP asset management modules [1]. This study addresses that gap by proposing a novel, BPR-driven framework that integrates bulk asset transfer mechanisms, optimized approval workflows, and forward-reverse system integration. The research contributes to both theory and practice by offering a scalable, adaptable solution that improves operational efficiency, strengthens data integrity, and reduces labor costs—demonstrating a significant advancement over existing approaches in ERP process optimization.

B. Asset Management Processes

Enterprise Resource Planning (ERP) is a comprehensive software solution that integrates various business processes such as finance, production, and sales—into a unified system. Within an ERP system, asset management refers to the structured approach of managing an organization's physical assets, including machinery, equipment, and vehicles, throughout its lifecycle. This encompasses key processes such as asset acquisition, operation, maintenance, and disposal to ensure optimal utilization and efficiency. Effective asset management in ERP systems is essential as it centralizes asset-related data and integrates it with other critical business functions, such as finance, procurement, and maintenance. This integration provides real-time visibility into asset performance, cost tracking, and maintenance schedules, enabling organizations to maximize asset utilization, reduce operational costs, and improve decision-making. Businesses that manage and track their assets effectively across multiple departments benefit from enhanced productivity, streamlined operations, and improved financial reporting [12].

Standard ERP applications often require customization to align with an organization's operational needs. Customized integration plays a crucial role in enhancing asset management workflow, approval processes, and data accuracy. While most ERP systems provide basic asset management functionalities, organizations requiring advanced asset tracking, real-time data synchronization, or customized reporting must implement custom modules, modify existing functionalities, or integrate third-party software [13]. Asset management is critical for large enterprises that manage extensive inventories across multiple locations. However, the literature highlights several challenges, including:

- Inefficient manual processes that slow down operations.
- Lack of real-time data synchronization, leading to inconsistencies in financial reporting.
- Complexity in managing large-scale asset transfers across multiple locations.

Modern asset management strategies increasingly prioritize sustainability by incorporating lifecycle cost management

and reliability-centered maintenance. Organizations now emphasize long-term cost reduction and environmental impact minimization through efficient asset utilization and responsible disposal strategies. The evolution of topic modeling in asset management research further highlights a shift towards advanced analytics, predictive maintenance, and AI-driven decision-making [14]. ERP automates key asset management processes such as asset transfers, additions, disposals, and reclassifications. These automated workflows minimize manual errors and enhance operational efficiency. The system seamlessly integrates with ERP applications such as payables, project costing, and ledger accounting, ensuring accurate data flow and financial reconciliation. Additionally, assets provide robust reporting tools, offering organizations greater visibility, control, and compliance over their assets throughout their lifecycle [15], [16].

Security is a ERP-based asset management. The literature underscores the importance of integrating security protocols into ERP systems to safeguard sensitive business data and operational integrity. A well-defined security framework ensures access control, compliance with industry regulations, and protection against cyber threats. Research highlights the necessity of embedding IT and corporate security measures into ERP systems to create a holistic, organization-wide security strategy. This approach enhances business process continuity, data governance, and system reliability across multiple operational domains, including technology, communication, and functionality [17].

Cloud-based ERP solutions have changed asset transfer and management by increasing scalability, accessibility, and real-time data analytics. In distributed enterprises, where assets are regularly moved between locations, cloud ERP provides centralized, automated, and intelligent solutions for tracking, validation, and compliance.

1) Scalability for large-volume asset related data: Traditional ERP systems often struggle to handle large-scale asset management related data owing to system limitations and manual processing. Cloud-based solutions, dynamically scale to accommodate high-volume asset related data across multiple locations. This prevents the challenges associated with traditional on-premises ERP setups, ensuring that businesses can quickly reallocate assets without disrupting their operations.

2) Accessibility for real-time asset tracking: One of the biggest challenges in asset management is ensuring that all stakeholders have real-time access to the asset transfer data. Cloud-based ERP solutions provide anytime and, anywhere visibility of asset locations, movement status, and transfer history. This is particularly beneficial for enterprises that, manage assets across multiple regions, warehouses, and business units.

3) Real-time data analysis for decision making: Cloud ERP systems utilize AI-driven analytics and machine learning to optimize asset transfers and management decisions. Organizations can predict asset availability based on past transfer patterns, identify idle assets for reassignment, and automate depreciation and cost adjustments during transfer.

C. BPR Methodologies

In Business Process Reengineering (BPR) and system analysis, the As-Is system refers to the current state of business processes and systems as they exist today. This includes the existing workflows, tools, and technologies utilized within an organization. The primary goal of As-Is analysis is to gain a comprehensive understanding of current operations, identifying inefficiencies, bottlenecks, and manual interventions that hinder performance. By analyzing the As-Is system, businesses can pinpoint areas for improvement, such as outdated processes, redundant tasks, and missing functionalities that lead to operational slowdowns and inconsistencies [18]. In contrast, the developed system represents the optimized future state of the current system after reengineering, and improvements have been implemented. It defines how processes should function once redesigned to eliminate inefficiencies, enhance automation, and optimize workflow. The framework of developed system focuses on integrating advanced technologies, automating repetitive tasks, and improving process accuracy, ultimately leading to cost reductions, enhanced scalability, and improved organizational efficiency. This future-state model serves as a roadmap for transformation, ensuring that the organization operates in a streamlined and sustainable manner [18].

In BPR, the As-Is and To-Be frameworks are essential for diagnosing inefficiencies and designing optimized workflows. The As-Is process documents current business processes, highlighting bottlenecks, redundancies, and areas for enhancement. For asset management, the As-Is state is often characterized by manual, error-prone processes that lack real-time synchronization. The To-Be framework, on the other hand, envisions an optimized asset management process where automation, system integration, and advanced security controls improve efficiency, accuracy, and data consistency. Pereira and Silva (2020) emphasized the importance of analyzing the As-Is state to fully understand data synchronization challenges and inefficiencies in asset transfer processes. Transitioning from As-Is to To-Be requires redesigning workflows, improving system interoperability, and implementing role-based access controls to secure asset data and enhance the operational efficiency. The literature provides Numerous case studies have demonstrated successful BPR-driven transitions from manual to automated asset management systems, resulting in significant improvements in cost savings and workflow efficiency.

Prior research presented a quantitative verification approach for optimizing design models using Business Process Model and Notation (BPMN) and transforming them into improved models [19]. It addresses timed constraints, probabilistic tasks, and dependencies between activities in the processes. This technique helps designers identify operational gaps, time issues, and flaws, thereby ensuring a more effective process design [20].

Several Business Process Reengineering (BPR) methodologies have been applied to asset management within ERP-integrated environments, each offering unique advantages in optimizing asset workflows, reducing inefficiencies, and enhancing data-driven decision-making. Among these methodologies, Business Process Maturity Modeling (BPMM) can be used to evaluate the As-Is asset transfer process and new asset transfer process in the developed system. Previous research has specified that the Business Process Maturity Model (BPMM) is a structured framework used to assess and improve the efficiency, standardization, and automation of business processes. Originating from process maturity models such as the Capa-

bility Maturity Model (CMM), BPMM provides a systematic approach to evaluating the current state of a business process and identifying areas for improvement. Categorizing processes into five distinct maturity levels—initial, Managed, Defined, Quantitatively Managed, and Optimizing—BPMM enables organizations to transition from ad hoc, manual processes to fully automated, data-driven decision-making systems [21].

Among the other methodologies, Lean focuses on eliminating waste and streamlining asset management processes by removing non-value-added activities such as redundant approvals, excessive manual data entry, and delays in asset transfers [22]. In ERP-integrated asset management, Lean principles enable faster asset tracking, optimized maintenance schedules, and reduced downtime by ensuring that resources are efficiently allocated [23].

Six sigma, a data-driven methodology, enhances ERP-integrated asset management by minimizing errors in asset tracking, valuation, and transfer processes [24]. Using the Define, Measure, Analyze, Improve, and Control (DMAIC) framework, Six Sigma ensures that asset data remains accurate, eliminating inconsistencies that could affect financial reporting and compliance. Organizations implementing Six Sigma in ERP-based asset management benefit from predictive maintenance models, optimized asset utilization, and reduced lifecycle costs [25].

More over, Kaizen, which emphasizes continuous improvement, plays a crucial role in ERP-integrated asset management by encouraging incremental process enhancements. Organizations adopting Kaizen-driven asset management continuously refine workflows, such as real-time asset tracking, proactive maintenance alerts, and automated asset reconciliation, ensuring that ERP systems evolve with business needs while reducing operational risks [26].

Rather than the above methodologies, Agile BPR is particularly effective for dynamic asset management environments where assets frequently move across locations or undergo rapid reallocation [27]. By implementing iterative process improvements, Agile BPR ensures that ERP-integrated asset workflows remain flexible, enabling faster asset transfers, improved collaboration among departments, and adaptive compliance frameworks. Agile methodologies also enhance cloud-based ERP integrations, where businesses need real-time analytics and automated decision-making for asset lifecycle management.

However, Total Quality Management (TQM) focuses on a comprehensive process control across asset management operations. When integrated into ERP systems, TQM ensures standardized asset tracking protocols, regulatory compliance, and high-quality data integrity across departments. TQM-driven ERP systems automate asset depreciation calculations, optimize audit trails, and enhance approval workflows, minimizing manual intervention and compliance risks [28].

Rather than the above methodologies, the t-test and Analysis of Variance (ANOVA) will be used to analyze the success metrics of the proposed system, which are statistical methodologies used to compare the means between groups. As in prior research, A T-test is typically used when comparing the means of two groups to determine if there is a significant difference between them [29]. This comes in two main forms: the independent t-test (for comparing two independent groups) and the paired t-test (for comparing two related groups, such as pre-test and post-test data). On the other hand, ANOVA is used when comparing means across three or more groups to assess whether at least one group differs significantly from the others. Unlike multiple t-tests, which increase the risk of Type I errors, ANOVA controls for this error rate and provides a more reliable comparison. The most common form is one-way ANOVA, which examines one independent variable, whereas two-way ANOVA analyzes two independent variables and their interactions. Both methodologies assume normality and homogeneity of variances which will be properly used in the research.

When applied to ERP-based asset management, each BPR methodology contributes to a more efficient, automated, and data-driven approach. While Lean and Agile BPR enhance flexibility and process adaptability, Six Sigma and TQM provide structured, quality-focused improvements. Kaizen ensures continuous refinement, making ERP systems more responsive to business changes. The hybrid approach of Lean Six Sigma is particularly effective in large-scale ERP implementations, ensuring both process efficiency and data accuracy in asset management operations.

D. Asset Management Technologies

The Asset Management Modules of ERP systems lack critical functionalities for efficient bulk asset transfers, creating significant operational inefficiencies. The current system requires manual, one-by-one asset transfers, making the process time-consuming, labor-intensive, and prone to errors, especially for organizations managing large-scale asset inventories. Additionally, the absence of automated approval workflows forces users to rely on email and manual approval, leading to delays and compliance risks. Another major limitation is the lack of real-time data synchronization, resulting in inconsistencies between transferred assets and system records, which affects financial reconciliation and decision-making.

Furthermore, the high licensing costs for user management make it expensive for organizations to assign and control user roles efficiently. ERP systems also lack a dedicated budgeting module for asset transfers, which makes financial planning and expenditure tracking challenging. Finally, the reporting and audit trail features in the current system are insufficient for organizations that require comprehensive tracking of asset movements and compliance monitoring.

This research addresses these critical gaps by developing the Oracle Asset Transfer Suite, an enhanced module designed to introduce bulk asset transfer capabilities, automated approval workflows, real-time data synchronization, cost-effective user management, budget tracking, and advanced reporting functionalities, ensuring efficiency, cost reduction, and improved data accuracy in Oracle ERP's asset management processes according to the use case we selected with the financial institute in Sri Lanka.

Research indicates that leveraging technologies such as Oracle APEX for application development, JavaScript for dynamic user interfaces, and PL/SQL for efficient database operations significantly improves workflow automation, data

accuracy, and operational efficiency [30]. Additionally, advanced analytics tools, such as Power BI, have been integrated into asset management systems to enhance decision-making, provide actionable insights, and optimize process performance.

III. Research Methodology

This study employs a mixed-method approach, integrating both qualitative and quantitative methods, to conduct a comprehensive analysis of asset management processes within Oracle ERP. The qualitative approach focuses on gathering insights from key stakeholders through interviews, observations, and process evaluations to identify inefficiencies, bottlenecks, and user challenges in the current As-Is system. The quantitative approach involves collecting and analyzing system performance data, including process times, error rates, and asset utilization statistics, to assess the effectiveness of both existing and developed system. By combining these two approaches, this study ensures a holistic understanding of the asset management process, capturing both subjective user experiences and objective performance metrics that inform the re-engineering of asset workflows. This methodology enables data-driven evaluation of process improvements, ensuring that the proposed solutions enhance efficiency, accuracy, and automation.

To conduct this study, a representative sampling method was employed. The total population considered in this study consisted of 256 branches distributed across nine provinces of Sri Lanka. A 20% sample was selected from this population to ensure a representative subset for analysis. The sample size was calculated using Eq. (1) below.

$$Samplesize = 20/100 * 256 = 51.2 \quad (1)$$

Because the sample size must be a whole number, it was rounded to 51 branches. Thus, the research findings and analysis were based on a sample of 51 branches drawn from the entire population. The selected 51 branches were distributed among the nine provinces as follows:-

- Western Province – 10 branches (21.28%)
- Eastern Province – 8 branches (17.02%)
- Northwestern Province – 7 branches (14.78%)
- Northern Province – 5 branches (10.64%)
- North Central Province – 5 branches (10.64%)
- Central Province – 4 branches (8.51%)
- Sabaragamuwa Province – 3 branches (6.38%)
- Southern Province – 3 branches (6.38%)
- Uva Province – 2 branches (4.26%)

This distribution ensures that all provinces are represented in the study, while maintaining proportionality relative to the total number of branches in each region. The sampling approach enhanced the validity and generalizability of the research findings by capturing variations across provinces.

A. Data Collection Methods

The researchers will rely on both primary and secondary data sources from a reputed financial institution in Sri Lanka to analyze the current asset management process of Oracle ERP and develop an optimized solution.

1) Primary data collection:

a) Direct observations: Systematic observations of current asset management workflows are conducted to identify inefficiencies, manual interventions, and synchronization issues in the As-Is process. This will provide real-world insights into operational challenges during asset transfers and system integration.

b) Interviews with key stakeholders: Semi-structured interviews will be conducted with approximately five key stakeholders, including asset managers, IT personnel, and finance teams, to gain qualitative insights into the limitations of the existing Oracle ERP system and expectations of the developed system. These interviews focused on pain points, process delays, security concerns, and system usability.

c) User surveys: A structured survey was distributed to approximately 20 Oracle ERP users within the institution. The survey collects quantitative data on system performance, user satisfaction, error rates, and perceived inefficiencies. The results will help validate the findings from the observations and interviews, while providing a broader perspective on system challenges.

2) Secondary data collection:

a) Historical data and system logs: An analysis of historical asset transfer logs, ERP performance reports, and system audit trails will be conducted to quantify existing inefficiencies. This will serve as a baseline for comparing As-Is and developed systems and help in designing a customized asset transfer module.

b) ERP documentation and regulatory guidelines: A review of system documentation, compliance standards, and financial regulations relevant to asset management will be conducted to ensure that the proposed reengineering aligns with the best practices and industry requirements.

This structured data collection strategy ensures a comprehensive evaluation of existing inefficiencies, providing the foundation for an optimized, automated, and secure asset management solution within Oracle ERP.

B. Process Mapping Techniques to Model the Business Processes

To systematically analyze and optimize the asset management workflow, this study employs flowchart, Business Process Model and Notation (BPMN) Diagrams, and Swimlane Diagrams to map both As-Is and improved processes. These tools provide a structured, visual representation of process flows, enabling the identification of inefficiencies, manual interventions, and areas for automation [31]. Flowcharts are used to visually represent the step-by-step sequence of the current asset management process. Each step illustrates the flow of actions, decision points, and manual interventions, providing a clear, high-level overview of the existing workflow. This will help pinpoint bottlenecks, repetitive tasks,

and redundant approvals that contribute to inefficiencies, such as delays in asset transfers or excessive manual data entry. Additionally, these flowcharts will serve as discussion aids during stakeholder interviews and surveys, facilitating the identification of pain points in the current system. Moreover, Business Process Model and Notation (BPMN) diagrams were employed to provide a detailed, standardized representation of both the As-Is and improved processes. The BPMN enables the modeling of complex workflows, incorporating elements such as: activities (tasks performed at each stage), decision points (approval or rejection conditions), and data flows (information exchange across departments). By leveraging the BPMN, this study documents the current inefficiencies in asset management workflows and identifies opportunities for process reengineering. In particular, BPMN diagrams of developed system highlight automation opportunities, such as automated approval workflows to reduce manual interventions, system-driven asset tracking for improved real-time synchronization, and enhanced Oracle ERP integration to eliminate redundant data entry. This structured mapping ensures that every step in the process is carefully analyzed, allowing for a well-defined transformation strategy. In addition, swimlane diagrams will be utilized to clearly define the roles and responsibilities of the various departments and stakeholders involved in asset management. By mapping the interactions between different entities, swimlane diagrams will help identify gaps in communication and workflow handovers, highlight role-based inefficiencies, such as duplicated approvals or unclear accountability, streamline the improved process by redistributing responsibilities and automating handoffs, and ensure a more structured and, efficient workflow for asset transfers.

By incorporating these process mapping tools, the study will create a comprehensive blueprint for optimizing asset management workflows within Oracle ERP, ensuring a data-driven, well-documented transition from the As-Is to the improved process state.

1) Criteria for analysis: The criteria listed in the Tables II, III, and IV will be used to evaluate the As-Is process and design the optimized To-Be process.

By following this methodology, this study comprehensively evaluates the current asset management process and proposes a reengineered solution that improves efficiency, reduces costs, and mitigates risks, while leveraging technology to achieve an optimized To-Be state.

TABLE II. SUMMARY OF PROBLEMS IN AS-IS PROCESS

Category	Asset Transfer Process is Too Long	High Manual Data Entry	Repeated Processes	Unnecessary Processes	Repetitive Tasks
Quantity				✓	
Eliminate	✓		✓		
Combine			✓		
Separate	✓				
Simplify	✓				
Automate	✓	✓			✓

TABLE III. SUMMARY OF PROBLEM AREAS IN AS-IS ASSET MANAGEMENT

Problem Area	Impact on Efficiency	Potential Solution	Priority Level
Asset transfer process is too long	Significant delays in operations	Automate bulk asset transfers	High
High manual data entry	Increases labor costs and errors	Automate data entry and validation	High
Repeated processes	Redundancies create inefficiencies	Streamline and eliminate redundant steps	Medium
Unnecessary processes	Time wasted on non-essential tasks	Remove unnecessary approval steps	Medium
Repetitive tasks	Employee dissatisfaction	Automate repetitive tasks	Low

TABLE IV. SUMMARY OF PROBLEM SOLUTIONS FOR TO-BE PROCESS

Problem	Proposed Automation	Efficiency Improvements	Priority Level
Asset transfer process is too long	Bulk asset transfer capability	Reduces process time by 50 percent	High
High manual data entry	Automated data input	Reduces data entry errors	High
Repeated processes	Workflow automation	Streamlines redundant processes	Medium

C. Case Study

For this research, a prominent financial organization with 256 branches and over 450+ Oracle ERP users in Sri Lanka was selected, which is recognized as a leading life insurance provider with solid financial standing and market presence. Operating as a subsidiary of a major Sri Lankan bank, the institution benefits from its parent comapny's financial strength and credibility. With continuous growth in revenue and market share, it has maintained a dominant position in the local finance sector. Listed on the Colombo Stock Exchange, the institution prides itself in a dedicated workforce, contributing to its sustained success and industrial leadership. The institution's vision emphasizes becoming Sri Lanka's most admired and trusted insurance partner, while its mission focuses on delivering innovative and, customized solutions with a passion for excellence and teamwork. Its commitment to ethical governance, customer-centricity, continuous improvement, and industry compliance reflects its dedication to exceeding stakeholder expectations.

Discussions with the institution, revealed that the manual and time-consuming process of transferring assets individually, highlighted inefficiencies, and the substantial amount of manual work involved. The core challenge was bulk asset transfer functionality. Consequently, the institution experienced delays in asset requests, assignments, and transfers, leading to considerable gaps in communication and approval processes. This situation presentes an opportunity for research, aimed at enhancing organizational efficiency through the application of Business Process Reengineering (BPR) principles, focusing on streamlining workflows and improving operational performance.

D. Ethic Statement

This study did not involve the use of any personal or sensitive data, as such information was not required for the scope of the research. The data collected were strictly limited to asset transfer records obtained from a relevant financial organization for the purpose of evaluating and enhancing the efficiency of the asset transfer process.

Permission to access and use these data was obtained through direct communication with the organization. Screenshots of the email correspondence granting this permission have been included and can be accessed via the following link.

https://drive.google.com/drive/folders/1T5xPlKda_ZW_GLIUhrgn_qO6HLVHi627

The study did not involve any minors, and therefore, no parental or guardian consent was required. Additionally, as this research is not medical in nature, no medical records or health-related data were used or accessed in any part of the study.

IV. Results and Discussion

A. As-Is System Analysis

1) Explanation on As-Is system: The current asset management process depicted in Fig. 1 in Oracle ERP is predominantly manual and involves several steps to request, transfer, and update asset records across various departments and locations.

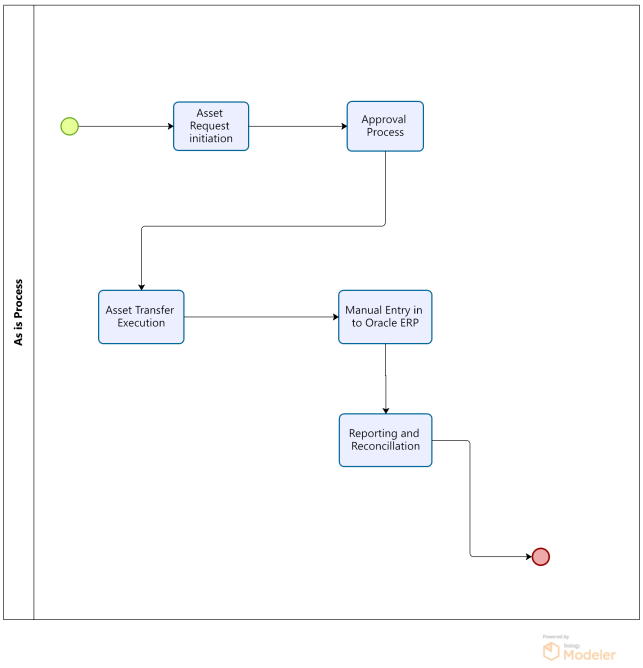


Fig. 1. Process of AS-IS System.

The current asset transfer process in Oracle ERP begins when a department or branch submits an asset request, that is currently handled manually. Each request must include detailed information, such as asset ID, quantity, and location. Following this request, the process moves to an approval workflow, where the asset request passes through multiple stages involving department heads, finance teams, and asset managers. This approval process often experiences delays due to manual handoffs between stakeholders. Once the request is approved, asset transfer is initiated, however, the current system allows only one asset to be transferred at a time, with each transfer requiring manual logging into the system, making it a time-consuming and error-prone task. After the physical transfer of the asset, data entry personnel update the asset's status, location, and transfer details in the Oracle ERP system. This step frequently involves verifying the transfer with physical inventory records and manually entering the data into various system modules. Finally, finance and audit teams must reconcile asset movements with financial records at regular intervals to ensure compliance with company policies and regulations.

The Business Process Maturity Model (BPMM) was utilized to assess the matuariry of the above as is process. The BPMM classifies process maturity into five levels: initial, managed, defined, quantitatively managed, and optimize. Each level represents the degree of process standardization, automation, and data-driven decision-making in an organization.

At Level 1 (Initial), the processes are unstructured, ad hoc, and reliant on individuals rather than defined workflows. The current asset transfer process in Oracle ERP aligns with this level, as it remains largely manual, time-consuming, and inefficient. To address these inefficiencies, automation is recommended, particularly for bulk asset transfers and validation mechanisms. This would reduce human dependency and improve the process reliability.

At Level 2 (Managed), some degree of process discipline is introduced, although inconsistencies persist. The existing asset transfer mechanism in Oracle ERP has seen partial standardization, but manual interventions still leads to delays. To advance to the next level, it is essential to establish standardized workflows and integrate automated approval mechanisms that eliminate the bottlenecks caused by human involvement.

Progressing to Level 3 (Defined) requires processes to be well-documented, standardized, and uniformly followed across the organization. The asset transfer module currently has partially defined workflows, but lacks automation. To enhance operational efficiency, a system-based request-tracking mechanism should be introduced to, ensure streamlined workflows and improved transparency in asset transfers.

At Level 4 (Quantitatively Managed), the processes are controlled using data-driven approaches. The current form of the asset transfer module does not support real-time monitoring or performance measurements using Key Performance Indicators (KPIs). The integration of dashboards and analytics tools enables organizations to track asset transfers effectively, monitor key metrics, and make informed decisions based on data.

Finally, at Level 5 (Optimizing), organizations focus on continuous process improvement through predictive analytics and feedback loops. The absence of such mechanisms in Oracle ERP's asset transfer module results in a lack of proactive decision-making and process enhancements. Implementing predictive analytics for asset lifecycle management would provide valuable insights, allowing organizations to anticipate

issues, optimize resource allocation, and enhance overall asset utilization.

This explained BPMM analysis is shown in the Table V.

TABLE V. BPMM (BUSINESS PROCESS MATURITY MODEL) FOR AS-IS PROCESS ANALYSIS

Maturity Level	Description	Current State	Target Improvements
Level 1 - Initial	Processes are ad-hoc, unstructured, and dependent on individuals	The current asset transfer process is manual, lengthy, and inefficient	Implement automation for bulk transfers and validation mechanisms
Level 2 - Managed	Basic process discipline exists, but inconsistencies remain	Some standardization in Oracle ERP, but manual interventions cause delays	Standardize workflows, introduce approval automation
Level 3 - Defined	Processes are well-documented, standardized, and followed organization-wide	Partially defined processes, but lack automation	Implement system-based request tracking and streamlined workflows
Level 4 - Quantitatively Managed	Processes are measured and controlled using data-driven approaches	No real-time monitoring, lack of KPIs	Introduce dashboards and analytics for asset tracking and performance measurement
Level 5 - Optimizing	Continuous process improvement through predictive analytics and feedback loops	No continuous feedback mechanism or improvements	Implement insights and predictive analytics for asset life cycle management

2) As-Is system process mapping: The As-Is process can be represented by the following Fig. 2 which highlights the steps involved from the asset request to reconciliation.

The asset transfer process in the provided diagram begins with branch logging into the company email and submitting an asset request. The procurement department, using the Oracle Fusion Cloud ERP, creates a purchase order based on the request. Once the assets are physically received, the inventory department generates a Good Receive Note (GRN) to confirm receipt and proceeds to create asset transfers in the system. A decision point follows, where it is determined whether the transfer is a bulk or individual asset transfer. If the transfer involves multiple assets, the system utilizes an in-built Oracle Fusion spreadsheet for bulk transfers, whereas individual assets are transferred individually in the case of non-bulk transfer. After the asset transfer is completed, a confirmation email is sent to relevant stakeholders, marking the completion of the process.

3) Current asset transfer process analysis:

a) Data collection: To conduct a comprehensive analysis of the current asset transfer process within the selected financial institution, data were collected using both primary and secondary Data Collection Methods, as outlined in the section "Data Collection Methods." The data collection process encompassed departments and branches across all nine provinces of Sri Lanka according to the sampling technique. This analysis did not involve personally identifiable information. Upon completion of data collection, 47 records were obtained covering all user groups related to the research.

The collected data comprises various aspects such as, "user experience level, frequency of requesting asset transfers, thoughts of the users regarding the communication during asset transfers, inefficiency in current process, need of real-time synchronization, importance of auditing and compliance". The collected data-set consists of categories relevant to the asset transfer process, with a focus on key ideological areas. The primary categories of the data collected were as follows:-

- Demographic Data
  - Demographic data refers to information regarding the length of experience of employees engaged in the asset transfer process, categorized as follows.
    - Less than one year
    - 0 to 5 years
    - 5 to 10 years
    - more than 10 years
- Quantitative Data
  - Experience Distribution - Numerical data representing the number of respondents within each experience category across different departments and branches.
  - Quantitative measures of communication and collaboration metrics - indicate the levels of agreement or disagreement regarding the effectiveness of communication between departments involved in asset management.
  - Asset Request Frequency - Frequency data capturing how often respondents request assets, categorized as daily, weekly, monthly, or rarely.
  - Security and Compliance Perceptions - Data reflecting respondents' views on the efficacy of automated features for asset audits and compliance assurance.
- Qualitative Data
  - Challenges in Asset Requests - Descriptive data highlighting common difficulties encountered in the asset request process, such as longer approval timelines and errors associated with manual data entry.
  - System Feature Feedback - A combination of qualitative and quantitative responses regarding desired enhancements in the asset management system, including automation, real-time synchronization, and compliance report generation capabilities.
  - Stakeholder Expectations - Data capturing overall expectations for improvements in asset management processes and system functionalities.
- Effectiveness Ratings and Collaboration Assessment
  - Communication Effectiveness - Respondents' evaluations of the effectiveness of interdepartmental communication within the existing system.
  - Collaboration Assessment - Responses measuring the extent to which the current system facilitates collaboration among teams involved in asset management.

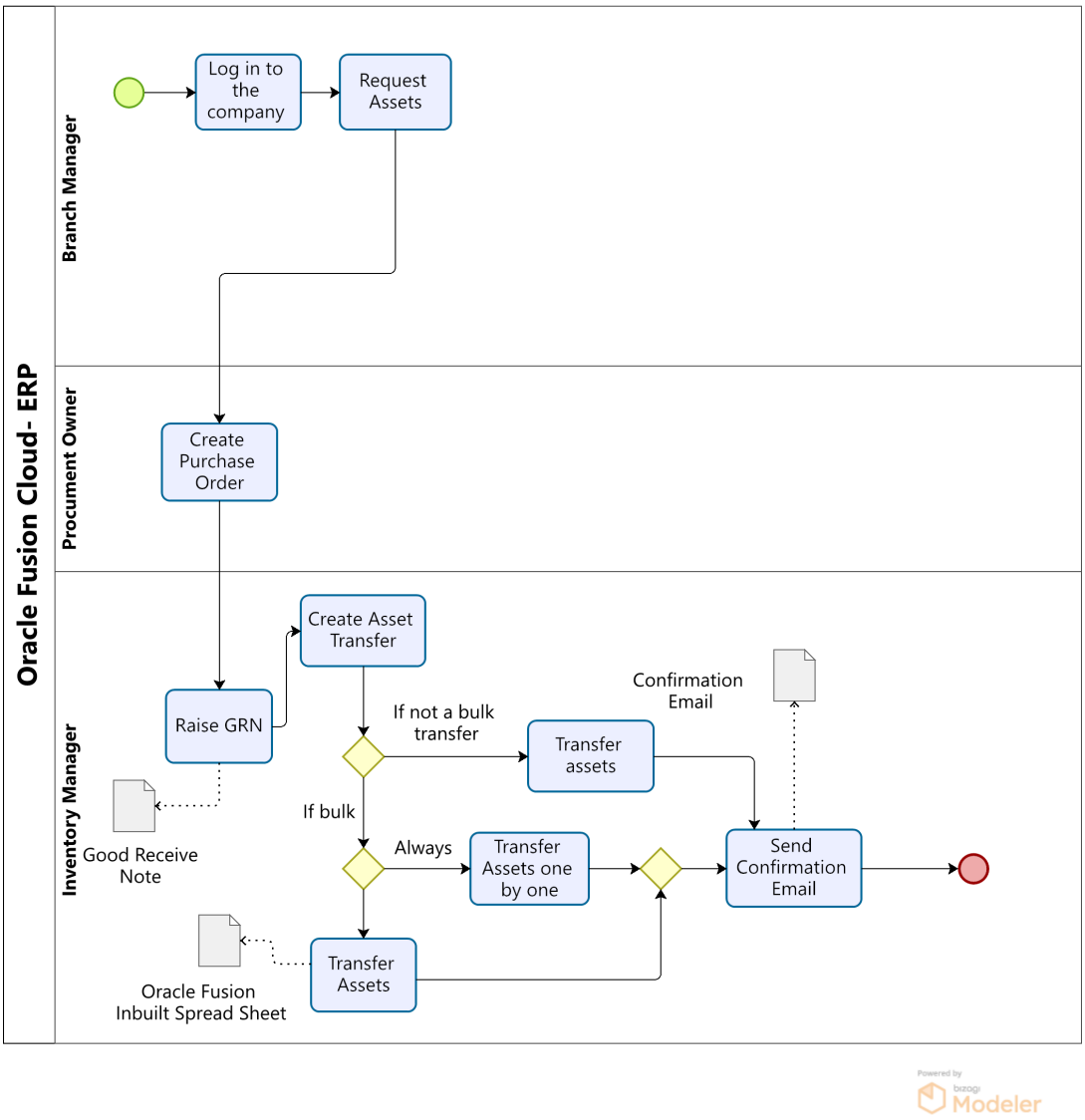


Fig. 2. Process mapping of As-Is system.

This structured approach to data collection enables a thorough understanding of the challenges, perceptions, and expectations surrounding the asset transfer process.

b) Data cleaning process: Before proceeding with the data analysis, a crucial preliminary step involved the data cleaning process to ensure the accuracy and reliability of the dataset. This process was essential for eliminating inconsistencies, errors, and irrelevant data, which could potentially affect the validity of the analysis.

After the data collection, the dataset was thoroughly examined to identify and remove invalid records. During this examination, various errors were detected, including repetitive records and incorrect data entries, which could distort the analytical outcomes. One of the key inconsistencies observed was the misplacement of the responses in different fields. Specifically, some users mistakenly entered the branch name in response to the province-related question, and, conversely, entered the province name in the branch-related question. Such misclassified entries compromised data integrity and required correction before the analysis.

To maintain data integrity, the following steps were undertaken:

- Removal of Duplicate Records – Identified and eliminated redundant records to prevent overrepresentation of certain data points.
- Elimination of invalid responses – Entries where responses were misclassified were removed, such as instances where users provided a branch name instead of a province or vice versa.
- Ensuring Data Consistency – Verified uniformity in categorical data, ensuring that each record was correctly aligned with the intended response category.

By addressing these inconsistencies, the dataset was refined to enhance its accuracy and reliability, ensuring that subsequent analysis would yield meaningful and valid results.

Upon completion of the data-cleaning process, 43 records were retained from the original 47. These refined records were subsequently utilized for the "As-Is system analysis", ensuring that only relevant and accurate data were considered for further evaluation.

c) Data analysis: Data analysis plays a critical role in understanding the efficiency and effectiveness of business processes within Oracle ERP. By analyzing key operational metrics, organizations can identify inefficiencies, smooth workflows, and improve employee satisfaction between branches and departments.

This analysis was conducted using Power BI, a powerful business analytics tool that enables data visualization and interactive reporting. Power BI allows users to connect with various data sources, efficiently process large datasets, and generate insights through dynamic dashboards and visual representations. By leveraging Power BI, the data were analyzed in a structured manner, making it easier to identify trends, patterns, and key areas for improvement in the bulk asset transfer process.

The As-Is system analysis was conducted to provide an in-depth analysis of the current state of the bulk asset transfer process within various departments and branches of the financial institution, highlighting the experience levels, bulk asset request frequencies, and critical importance of real-time synchronization and automation for future improvements. The analysis also highlights the key challenges faced by the existing systems.

According to the Fig. 3, 4, 5, 6, and 7, when considering the bulk asset request frequency by the departments and branches of the financial institution, the breakdown of the results is as follows. In addition, the analysis found that a significant proportion of employees involved in asset transfers fell under the "0 to 5 years" experience category. .

- Daily: 20.83%
- Weekly: 33.33%
- Monthly: 12.5%
- Rarely: 33.33%

The notably higher frequency of responses indicating "Rarely" can be attributed to the geographical distribution of the selected financial institutions. Most of their departments and branches are located in rural areas, where asset transfers occur infrequently. However, this trend shifts in urban regions, where demand for asset transfers is significantly higher, often occurring on a daily or weekly basis.

The high prevalence of daily and weekly requests highlights the ongoing operational need for an efficient bulk asset transfer process, particularly in high-activity departments where asset allocation is frequent. Notably, 33.33% of respondents rated real-time synchronization between Oracle ERP and other systems as extremely important. The responses varied regarding the effectiveness of communication between departments, with 39.53% indicating a positive experience. A strong 95.65% on the necessity for automated features in bulk asset transfers indicates a clear demand for reduced manual errors and increased operational efficiency. The key pain points indicated by branch/departmental users were, "Lengthy approval processes, manual data entry errors, and difficulty in tracking assets".

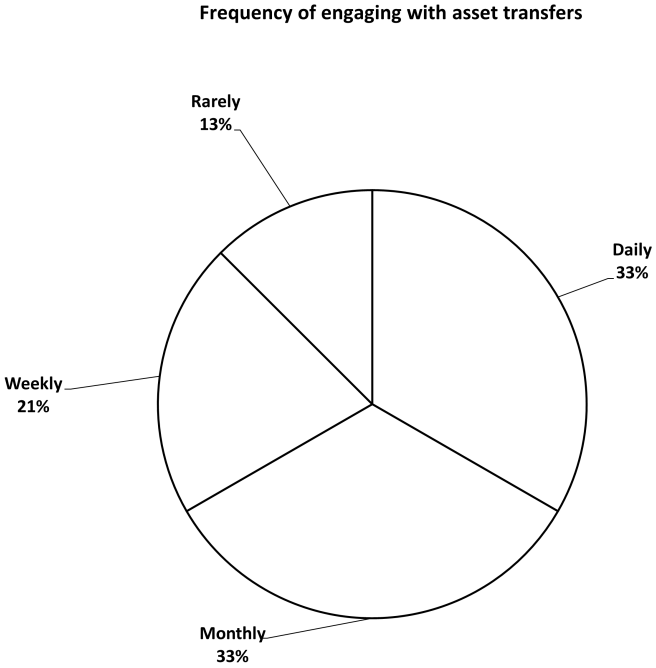


Fig. 3. Frequency of engaging with asset transfers from each branch/ department.

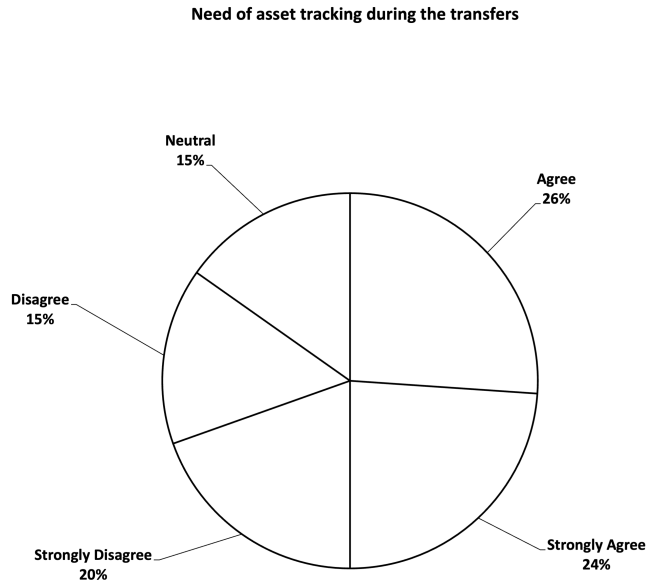


Fig. 4. Need of asset tracking during the asset transfers.

4) Identification of issues: The current asset management process suffers from several inefficiencies and bottlenecks.

- Lack of Bulk Asset Transfer The system does not support bulk transfers, forcing users to handle each asset individually. This leads to an increased time consumption and a higher chance of human error.

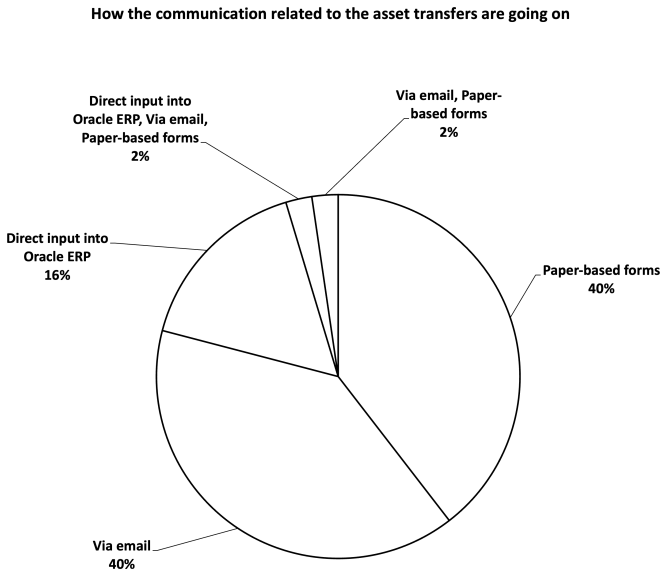


Fig. 5. How the communication related to the asset transfers are going on.

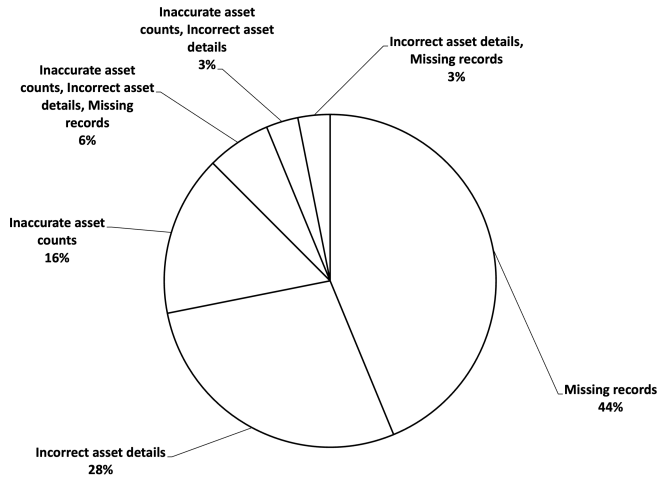


Fig. 6. Types of errors happen during the manual entries.

- The approval process requires multiple handoffs and manual interventions, which often lead to delays, particularly when approval is dependent on busy department heads or finance teams.
- Because asset data are manually entered into the system after physical transfers, there is a significant risk of errors in asset IDs, locations, and quantities. These errors can create discrepancies in financial reporting and asset tracking.
- Owing to delays in manual data entry and the lack of real-time updates, asset records across different departments and locations can become outdated, leading to inaccurate inventory counts.
- The manual nature of data entry and approval increases the risk of non-compliance with internal and external policies, particularly in terms of timely and accurate record-keeping.

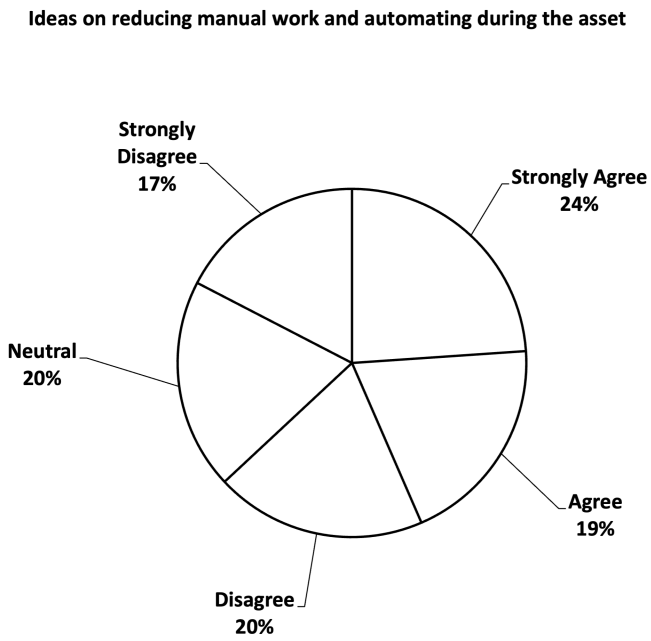


Fig. 7. Ideas on reducing manual work and automating during the asset transfer.

B. Design of the Developed System

1) Reengineering objectives: The objectives of reengineering the asset management process within the Oracle ERP system were,

Streamlining the asset transfer process by introducing bulk asset transfer capabilities can significantly improve operational efficiency. By reducing the time and effort required to move assets across departments or locations, businesses can enhance productivity. This approach also minimizes bottlenecks and allows staff to focus on more strategic tasks, rather than repetitive manual processes [32].

Automating key steps in the asset management process, such as- data entry, approval workflows, and error correction, can significantly reduce labor costs. Manual processes are time-consuming and prone to human error, both of which can lead to costly inefficiencies. By automating these tasks, businesses can not only save labor expenses but also increase the overall accuracy and speed of their asset management operations [33].

Strengthening compliance by improving data accuracy and real-time synchronization across departments is another crucial benefit. Discrepancies in asset data can lead to regulatory non compliance or internal auditing issues, which may result in fines or reputational damage. By ensuring real-time updates and synchronization, businesses can minimize these risks and, maintain accurate records for both regulatory and internal purposes [34].

Access controls are essential for ensuring the security, integrity, and compliance by regulating who can view, modify, or

execute specific actions within a system. These controls define the user roles, permissions, and authentication mechanisms to prevent unauthorized access, safeguard sensitive data, and enforce accountability. In enterprise systems such as Oracle ERP, access controls play a critical role in the above mentioned key steps in automation [35].

However, automation also plays a significant role in increasing data accuracy. Manual data entry is often riddled with errors,which can cause discrepancies in asset tracking and management. By automating data entry and synchronization processes, businesses can ensure that all asset movements are accurately reflected in the system, eliminating the potential for human error, and ensuring data integrity [36].

Finally, improving user experience through user friendly interfaces and minimizing manual interventions is essential for enhancing productivity and satisfaction. When users can interact with systems more intuitively, they are more likely to adopt technology and perform tasks more efficiently, leading to overall improvements in asset management [37].

2) Development of reengineered system: The Asset Transfer Suite is an application developed using Oracle APEX. It was created to offer a friendly scalable security platform for handling very large asset movements integrated directly with Oracle Fusion. They opted for Oracle APEX because it offered low-code architecture, rapid development, and Oracle database integration, thus providing quick deployment without performance or scalability being compromised. Architecture-wise, the application is designed to offer modularity and flexibility, with the business logic implemented in PL/SQL-encompassing critical activities like asset data validation, transformation, exception handling, and transaction rollback. Custom database objects such as tables, views, and packages were created to manage assets, transfer logs, audit trails, and user access controls.

Integration with Oracle Fusion ERP existed through both the web services using SOAP and REST. Web service callouts in Oracle APEX can be made through the use of a web source module or through PL/SQL's UTL_HTTP package. For authentication, if the particular Fusion API allowed OAuth 2.0 tokens, then such tokens were used; otherwise, basic authentication was used. Likewise, the integrations executed a series of activities like fetching the asset records, sending transfer requests, verifying transfer requests, and finally updating statuses of such requests, including updates regarding errors or validation messages issued by Fusion. The front-end interface of the application was built in APEX using its low-code UI builder, which allows the creation of interactive forms, reports, and dashboards in a snap. The users would filter assets by location, department, or cost center and perform those bulk transfers all at once, keeping track of their transaction history alongside real-time statuses synced with Oracle Fusion ERP.

Security integration was short and sweet-from Oracle Identity Cloud to Role-Based Access as defined within Fusion. Audit logs for transfer records, including user actions and system events, were created automatically for later traceability. Error handling and rollback mechanisms were designed and implemented to protect the data from accidental corruption in unplanned failure conditions. The entire suite was deployed and tested on staging Oracle Fusion test tenants. Functional testing, integration testing, and UAT provided a robust system that comprised the business requirements. After all validations were complete, the app was pushed into production. Later on, monitoring dashboards were added to keep track of API call success and failure rates so that issues could be addressed before anyone even noticed. This development approach gave the full capability of delivering a comprehensive, reliable, enterprise-grade solution that extended Oracle Fusion ERP for bulk asset transfers in an efficient and secure manner.

3) Developed system for reengineered process: The reengineering of the asset management process with adherence to the Oracle technology use guidelines, involves several transformative improvements, aimed at increasing efficiency, data accuracy, and process scalability. The key changes introduced include bulk asset transfer capability, automated approval workflows, real-time data synchronization, role realignment, and technology integration. These enhancements address the inefficiencies in the current system and offer a robust framework for future growth.

One of the most significant improvements is the introduction of bulk asset transfer capabilities. In the current system, assets must be transferred one at a time, which is time consuming and prone to delays. This new feature allows multiple assets to be simultaneously transferred, thereby accelerating the entire process. Bulk transfers significantly reduce the administrative burden, freeing up time for employees to focus on other critical tasks. This change is especially beneficial for organizations managing a large volume of assets, as it minimizes the number of individual transactions required and reduces repetitive tasks. By automating and simplifying asset transfers, organizations can expect faster processing times, reduced errors, and increased operational efficiency.

Another key feature of the reengineering effort is the introduction of automated approval workflow. In the current system, asset-related approvals often involve manual handoffs, that can result in bottlenecks and delays. The new automated workflow route asset requests to the appropriate stakeholders based on predefined business rules, ensuring that approvals are processed in a timely manner. This change reduces the need for manual intervention and eliminates the risk of requests becoming stuck or lost in the system. Automated workflows can also be customized to accommodate different approval levels based on the value or sensitivity of the asset, ensuring that all approvals are aligned with the organization's policies and procedures. By automating this process, companies can reduce turnaround times for asset requests, improve accountability, and enhance the overall transparency in asset management.

Real-time data synchronization is another crucial component of the asset management reengineering process. In many organizations, asset data are stored in disparate systems, leading to inconsistencies and outdated information. With the implementation of real-time synchronization, asset data are updated across all departments and locations instantly using web services. This ensures that all stakeholders have access to the most current and accurate data,thereby reducing the risk of errors caused by outdated information. Real-time synchronization is especially important for organizations that operate in multiple locations or have decentralized teams because it ensures that everyone is working with the same data. This capability also improves reporting accuracy and

enables better decision-making based on reliable, up-to-date asset information.

The reengineering effort also involves role realignment within the asset-management process. This change clarifies and automates the roles related to data entry, approval, and reporting. By separating these responsibilities, the organization can improve compliance and security and, ensure that tasks are handled by appropriate personnel. Access control is a key focus of this realignment, As it ensures that users have access to the information and functions necessary for their roles. This change enhances the overall security of the asset management process by reducing the risk of unauthorized access and improving accountability. Clear role definitions and automated workflows also help reduce human errors, as employees can focus on their specific tasks without having to navigate unnecessary complexity.

Finally, technological integration plays a vital role in the reengineered asset management process. The integration of advanced tools, such as Oracle APEX, for rapid application development allows the organization to build custom solutions tailored to its specific needs [38]. PL/SQL manages database operations, ensuring efficient and reliable handling of asset data, whereas JavaScript and HTML5 will be used to develop user-friendly interfaces that improve the usability of the asset management module [39]. These technologies provide a scalable platform that can grow with the organization, offering flexibility and adaptability for future needs. By leveraging these modern tools, the asset management system becomes more powerful and easier to use, thereby enhancing both user satisfaction and system performance. Moreover, the following results can be expected from the developed system, which can be identified using the BPMM methodology.

At Level 1 (Initial), the focus is on basic data retrieval and asset onboarding. By leveraging the Oracle SQL Cloud Developer and Oracle Autonomous Database, real-time asset data extraction can be ensured for timely updates. Additionally, the automation of asset onboarding and budgeting through Oracle APEX and Oracle Cloud Infrastructure eliminates manual asset entry while maintaining financial tracking accuracy.

Moving to Level 2 (Managed), the system setup and configuration phase ensures adaptability to business rules and policies using Oracle APEX. Furthermore, asset request and assignment processes are optimized using Oracle APEX and Oracle SQL Cloud Developer, leading to streamlined workflows, improved transparency, and reduced processing delays.

At Level 3 (Defined), bulk asset transfers are automated using Oracle APEX and Oracle Cloud Infrastructure, ensuring error-free and efficient transactions. Additionally, approval workflows are enhanced through APEX Workflows and APEX Email API, reducing manual interventions and expediting the approval processes.

In Level 4 (Quantitatively Managed), reverse integration with Oracle ERP enables seamless synchronization of asset updates, financial reconciliations, and status changes using Oracle SQL Cloud Developer and Oracle Autonomous Database. Performance tracking and analytics are further enhanced with dashboards that provide real-time monitoring and KPI tracking via the Oracle APEX and Oracle Cloud Infrastructure.

Finally, Level 5 (Optimizing focuses on strengthening security and compliance), Access control management is implemented using Oracle APEX and Oracle Identity Cloud Service (IDCS) to restrict unauthorized access and ensure compliance with security policies. Additionally, audit logging and compliance tracking, enabled by Oracle APEX Logging and Oracle Audit Vault, provide enhanced security, regulatory adherence, and visibility for asset transfers and approvals.

A summarized expected analysis from the improved process is shown in Table VI.

4) Modeling the developed system: The developed system for asset management can be visualized as shown in Fig. 8 and Fig. 9.

5) Key improvements in the developed system: The developed system introduces several key improvements that enhance the efficiency and streamline operations within the asset management system. First, the automated routing of asset requests and approvals based on predefined rules ensures that requests are directed to appropriate stakeholders without delay. This accelerates the approval process and minimizes the potential for errors, ensuring that all requests are handled consistently according to organizational policies. Additionally, the bulk asset transfer feature allows users to move multiple assets in a single operation, thereby significantly reducing the time and effort required for asset management tasks. This capability is particularly especially beneficial during audits, relocations, or system migrations, where managing large quantities of assets can be cumbersome.

Moreover, real-time automatic updating of asset records across the ERP system eliminates the need for manual data entry, greatly reducing the risk of inaccuracies. As changes are made, all related records are updated instantly, providing users with the most current information available. This integration enhances visibility into asset status and availability, allowing for better decision-making and resource allocation. Collectively, these improvements create a more responsive and reliable asset management process, empowering organizations to operate more efficiently and effectively.

6) Implementation strategy: To ensure successful implementation of the reengineered asset transfer process, a robust change management strategy was employed. This begins with stakeholder engagement, where key figures, such as asset managers, finance teams, and IT personnel are involved early in the process. Their insights were invaluable in addressing concerns and fostering a sense of ownership over changes. In addition, a clear communication plan was established to articulate the objectives, benefits, and specific changes associated with the new process, ensuring that all users understand the rationale behind the transition and are motivated to adopt it. Comprehensive training sessions were organized to equip users with the skills necessary to navigate new functionalities by user manuals and online resources to provide ongoing assistance.

Before full-scale implementation, user acceptance testing (UAT) was conducted in collaboration with the selected financial institution. This critical phase allowed end users to interact with the re-engineered system in a controlled environment, validating its functionality, usability, and alignment with operational requirements. UAT provided an opportunity to identify potential discrepancies, gather user feedback, and

TABLE VI. BPMM (BUSINESS PROCESS MATURITY MODEL) FOR DEVELOPED SYSTEM ANALYSIS - WITH ORACLE APEX ENHANCEMENTS

Maturity Level	Component	Description	Technology Used	Expected Outcome
Level 1 - Initial	Data Retrieval from Oracle ERP	Extract asset data from Oracle ERP for real-time updates	Oracle SQL Cloud Developer, Oracle Autonomous Database	Accurate and timely asset data retrieval with automated fetch mechanisms
	Asset Onboarding and Budgeting	Automate asset addition and integrate with budgeting controls	Oracle APEX, Oracle Cloud Infrastructure	Eliminate manual asset entry, ensure financial tracking with automated validations
Level 2 - Managed	System Setup and Configuration	Define configurable parameters for asset management	Oracle APEX	Flexible system that adapts to changing business rules and policies
	Asset Request and Assignment	Enable request submission and automated asset allocation	Oracle APEX, Oracle SQL Cloud Developer	Streamlined request processing, increased transparency, reduced delays
Level 3 - Defined	Bulk Asset Transfers	Automate and validate mass asset transfers within Oracle ERP	Oracle APEX, Oracle Cloud Infrastructure	Efficient, error-free bulk transfers with built-in approval workflows
	Approval Workflow and Notifications	Implement system-driven approval tracking and email notifications	Oracle APEX Workflows, APEX Email API	Faster approvals, role-based notifications, and reduced manual interventions
Level 4 - Quantitatively Managed	Reverse Integration to Oracle ERP	Sync asset updates, status changes, and financial reconciliations	Oracle SQL Cloud Developer, Oracle Autonomous Database	Real-time data updates in Oracle ERP with minimal manual intervention
	Performance Tracking and Analytics	Implement dashboards with KPIs and real-time monitoring	Oracle APEX, Oracle Cloud Infrastructure	Real-time asset tracking, performance measurement, and compliance reporting
Level 5 - Optimizing	Security and Access Control Management	Implement Access Control Lists (ACL), IP-based restrictions, and role-based permissions	Oracle APEX, Oracle Identity Cloud Service (IDCS)	Restricted system access, preventing unauthorized entry and ensuring compliance with security policies

refine the system to enhance its effectiveness. By involving end users in this testing phase, the organization could ensure that the final implementation meets practical expectations and minimizes disruptions during deployment.

Following a successful UAT, the technology deployment phase was commenced with a pilot testing stage, allowing the reengineered process to be trialed in a limited capacity across select departments or locations. This approach enabled the organization to gather further feedback and address any residual issues before a full-scale rollout. By adjusting based on both UAT and pilot outcomes, the organization could mitigate risks and ensure a smoother transition. Once the pilot proves successful, the full system was deployed organization-wide, maximizing the benefits of the new processes across all areas of operation.

Post-implementation, continuous monitoring and support were critical to the long-term success of the re-engineering process. A dedicated support team was established to address any challenges that may arise during the transition, ensuring that users had the assistance necessary to adapt effectively. Regular performance evaluations were conducted to assess system effectiveness and user satisfaction, allowing for ongoing improvements. By integrating user acceptance testing, structured deployment, and continuous monitoring, this comprehensive approach to change management enhanced operational efficiency while fostering a culture of collaboration and adaptability within the organization.

C. Comparison of As-Is and Developed Systems

1) Evaluation criteria: A comparison of the As-Is and developed asset management systems highlights significant improvements across various criteria. In terms of process efficiency based on time, below Table VII shows the how the process has optimized from the reengineered system.

When evaluating cost-effectiveness, the annual operational cost for the As-Is system is stand approximately $7,500 per user, which includes labor and error correction. With the implementation of the new system, these costs were not included, leading to a savings of $7,500 from users or a 100% reduction [40].

The compliance aspect reveals that the As-Is system experiences approximately 10 compliance incidents per year. The improved system reduced these incidents to only two, representing an impressive 80% improvement in compliance.

In terms of risk mitigation, the As-Is system has an error rate of 15% in asset transactions. The developed system has a significant lower error rate of 5%, resulting in a risk reduction of 67%.

Finally, in terms of asset utilization, the As-Is system achieves an asset utilization rate of 75%. The developed system has affected to enhance this rate to 90%, indicating a 20% improvement in how effectively the organization tracks, transfers, and allocates assets.

2) Analysis of changes: Tables VIII, IX, X, XI, and XII, depict the changes that performed in the developed system compared to As-Is System.

3) Impact on stakeholders: Tables XIII, XIV, and XV depict the impact of the developed System on stakeholders.

D. Quantitative and Qualitative Benefits

1) Quantitative benefits: Implementing automation in asset management can lead to significant cost savings, with an estimated 100% reduction in operational expenses by lowering labor costs and minimizing errors and reconciliation expenses. Additionally, the introduction of bulk asset transfers and automated workflows is projected to reduce the time required

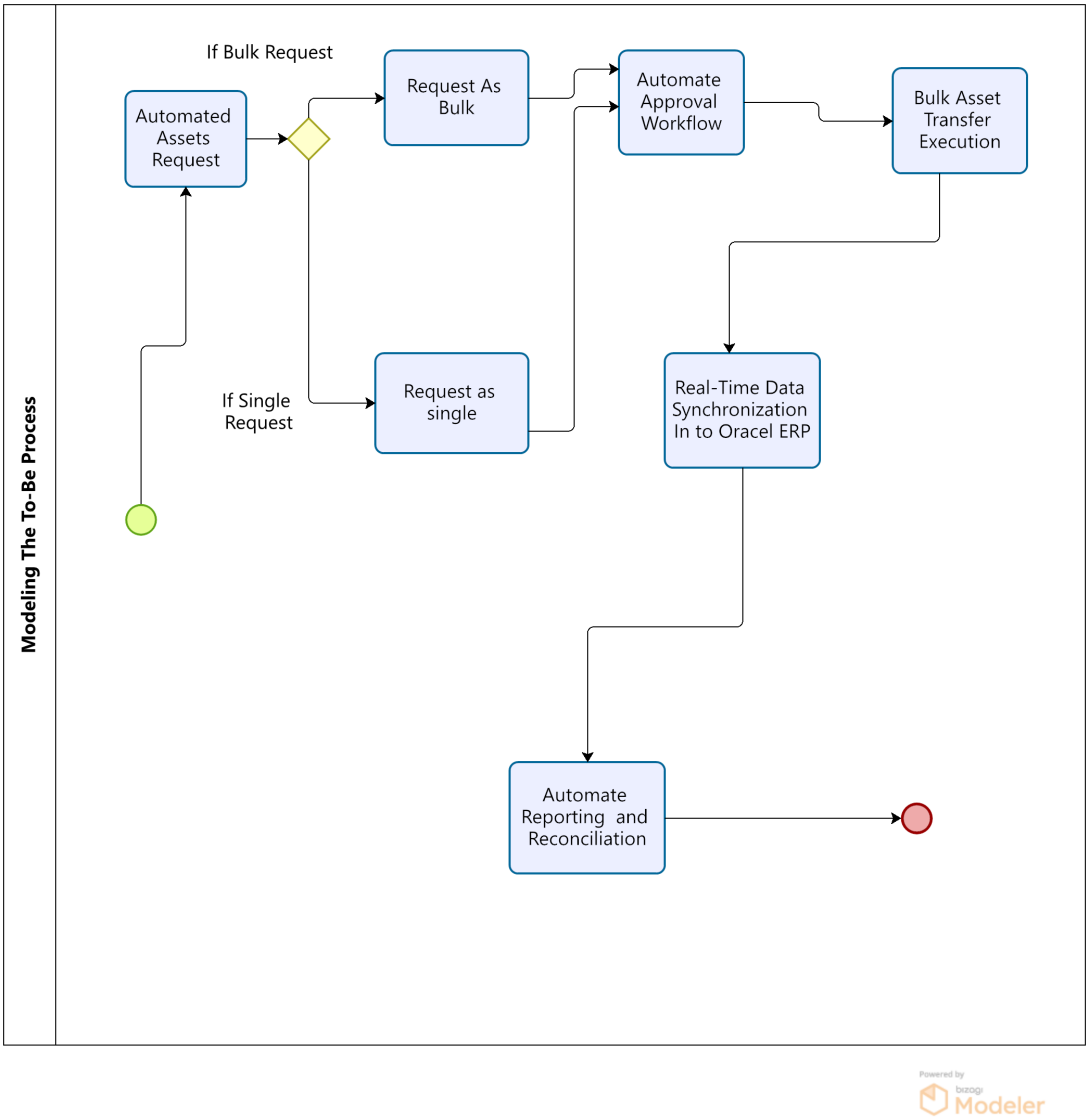


Fig. 8. Modeling developed system.

TABLE VII. COMPARISON OF THE PROCESS EFFICIENCY IN AS-IS SYSTEM AND REENGINEERED SYSTEM

Process Step	As-Is System (Oracle Fusion ERP Only)	Reengineered System	Time Saved
Time to Request Assets	16 hours	1 hour	15 hours
Time to Transfer Assets	48 hours	8 hours	40 hours
Time to Confirm notification of Assets	24 hours	2 hours	22 hours
Total Time for Full Process	88 hours (approx. 5.5 days)	11 hours (less than 2 days)	77 hours

for these transfers by 50%, thereby enhancing asset utilization and overall operational efficiency. Furthermore, automating data entry and ensuring real-time synchronization will decrease the error rate in asset transfers and record keeping, thereby reducing the need for expensive reconciliations.

2) Qualitative benefits: Transitioning from manual processes to automation is expected to significantly enhance employee satisfaction, as team members spend less time on repetitive tasks and more meaningful work, fostering a more positive work environment. Additionally, access to accurate, real-time data will empower management to make well-informed decisions regarding asset utilization, procurement, and financial planning, leading to more strategic asset management and improved resource allocation. Furthermore, with

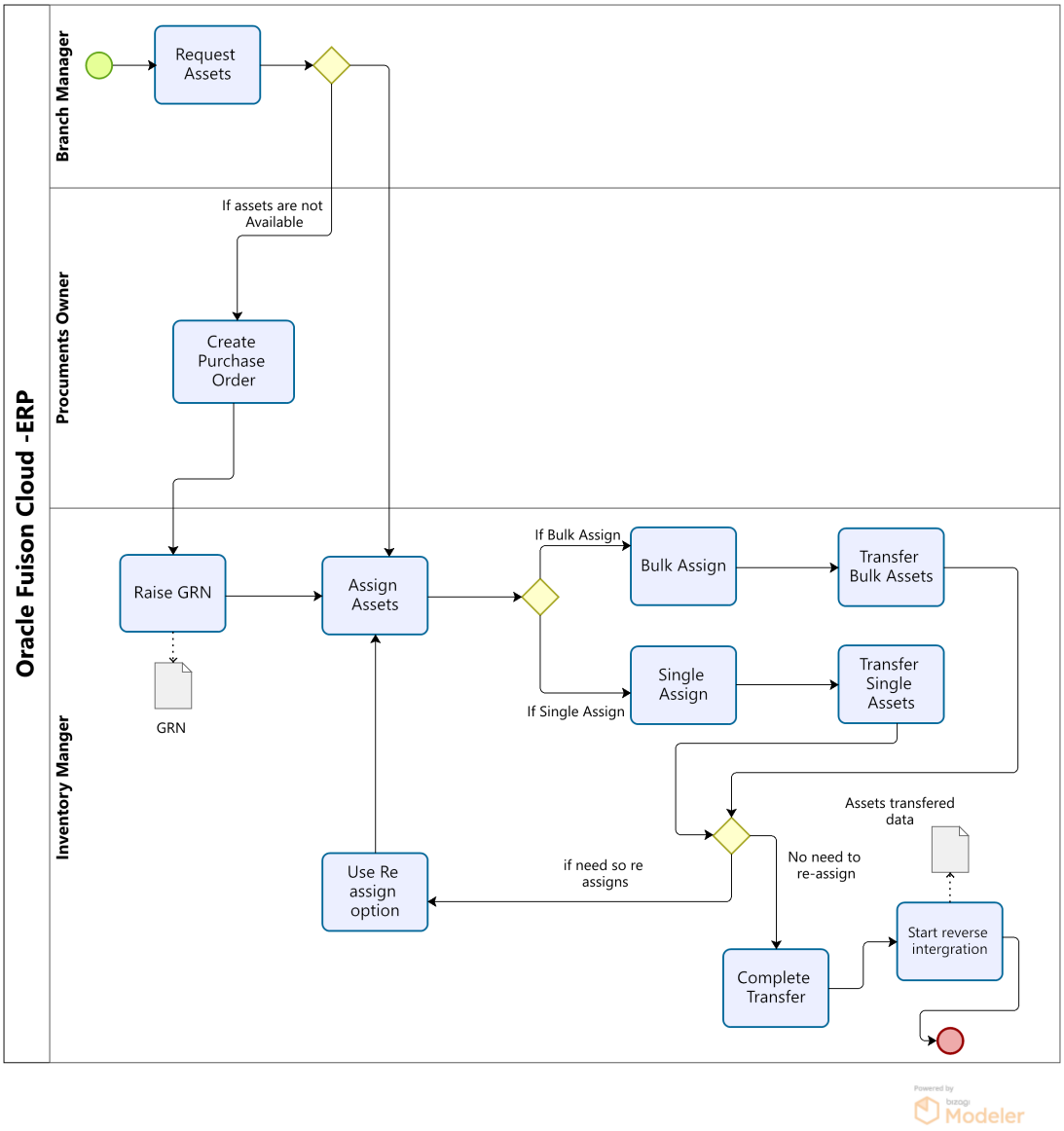


Fig. 9. Detailed developed system model.

TABLE VIII. PROCESS EFFICIENCY

As-Is System	Developed System
The current system requires manual input for each asset transfer, leading to significant delays, especially for large-scale transfers. The approval process is manual, involving multiple handoffs, which further slows down operations.	The bulk asset transfer feature in the developed system allows for transferring multiple assets at once, reducing the time required by up to 50%. Automated approval workflows eliminate manual handoffs, significantly improving the speed of asset requests and approvals.

TABLE IX. COST-EFFECTIVENESS

As-Is System	Developed System
High labor costs due to the manual nature of asset transfers, approvals, and data entry. Errors in manual data entry led to further costs in reconciliation and error correction.	Automation reduces labor costs by 30%, as fewer hours are needed for manual tasks such as data entry and error correction. The system also minimizes costs related to errors and operational delays.

improved compliance features and a lower risk of errors, the organization will be better positioned to meet both internal and external audit requirements, enhance its reputation, and minimize the risk of penalties.

In conclusion, the developed system offers significant improvements over the As-Is system in terms of process efficiency, cost-effectiveness, compliance, risk mitigation, and asset utilization. These changes are expected to have a positive impact on all stakeholders, delivering both tangible cost savings and intangible benefits such as improved satisfaction and decision-making capabilities.

TABLE X. COMPLIANCE

As-Is System	Developed System
The lack of real-time data synchronization leads to inaccurate records, making it difficult to comply with internal and external audit requirements. Delayed updates can result in financial reporting discrepancies.	Real-time synchronization of asset data ensures that records are always accurate and up to date, improving compliance with audit requirements. Automated workflows reduce the risk of non-compliance by ensuring that all approvals and asset transfers are properly documented.

TABLE XI. RISK MITIGATION

As-Is System	Developed System
The manual nature of the system introduces a high risk of errors in data entry, asset tracking, and approvals. Delays in asset transfers can cause operational risks, such as inventory shortages or unaccounted assets.	Automated data entry and real-time updates significantly reduce the risk of errors. The system's built-in validation checks during asset transfers further mitigate the risk of discrepancies. Timely and accurate record-keeping lowers operational and financial risks.

TABLE XII. ASSET UTILIZATION

As-Is System	Developed System
Poor visibility into asset locations and transfer statuses due to delayed and inaccurate data updates. This leads to under-utilization of assets and inefficiencies in resource allocation.	With real-time tracking and updates, asset visibility improves dramatically. The organization can better allocate resources and respond quickly to asset transfer requests, leading to improved asset utilization.

TABLE XIII. EMPLOYEES

As-Is System	Developed System
Employees, especially asset managers and data entry personnel, are burdened by repetitive manual tasks, leading to lower job satisfaction and higher error rates.	Employees, especially asset managers and data entry personnel, are burdened by repetitive manual tasks, leading to lower job satisfaction and higher error rates.

TABLE XIV. MANAGEMENT

As-Is System	Developed System
Management struggles with delayed reporting and inaccurate data, which hampers decision-making and makes it difficult to track asset utilization and costs.	With real-time data and automated reporting, management can make more informed decisions based on accurate and timely information. The improved visibility into asset movements and costs also enhances strategic planning and operational efficiency.

TABLE XV. EXTERNAL PARTIES (AUDITORS, REGULATORS)

As-Is System	Developed System
External parties often face difficulties auditing asset records due to discrepancies and outdated information, which can lead to compliance issues.	The Developed system ensures that external auditors and regulators have access to accurate and up-to-date records, improving compliance and reducing the risk of regulatory penalties.

V. CHALLENGES AND LIMITATIONS

A. Challenges in Reengineering

Resistance to change is a significant challenge in Business Process Reengineering (BPR). Employees often hesitate to adopt new technologies or processes because fear of job displacement and unfamiliarity with automated workflows. To effectively address this issue, organizations must implement clear communication strategies and comprehensive training programs.

Data limitations can complicate the BPR process. The accuracy and completeness of existing data in the As-Is system can be inconsistent or outdated, making it difficult to map current processes and identify inefficiencies. Resolving these data issues is crucial before transitioning to a improved system, where real-time data accuracy is essential.

Technological constraints also present another challenge. The integration of newly developed processes, such as bulk asset transfers and real-time data synchronization, requires an advanced technological infrastructure. Existing ERP systems may not support these changes without significant customization or upgrades, which can lead to compatibility issues and increased technical debt.

Finally, the implementation cost is a major concern. While the developed system is expected to reduce operational costs in the long run, the initial investment required for technology upgrades, training, and system deployment can create a financial burden for many organizations.

B. Limitations of the Research

The scope of this study is specifically centered on the reengineering of the asset management process within the Oracle ERP system. This focus inherently limits the generalizability of the findings, meaning that the insights and recommendations derived from this research may not be applicable to other ERP Systems or Asset Management platform. By concentrating on a single system, this study aimed to provide a detailed and nuanced understanding of the challenges and opportunities present within the Oracle ERP framework. However, this focus also means that organizations using different systems may not find the same relevance in the conclusions drawn.

Furthermore, the research relies heavily on existing data extracted from the current As-Is system. While these data serve as the foundation for process analysis, it is crucial to acknowledge that such information may be prone to inaccuracies or reflect incomplete records. These data limitations could significantly undermine the integrity of the process analysis, because the insights generated are only as reliable as the data on which they are based. If the existing data fail to capture the true state of asset management practices, the recommendations formulated in response may not effectively address the underlying issues or may overlook critical areas for improvement.

In addition to these internal limitations, this study must contend with various external factors that could impact the feasibility and effectiveness of the improved process. Factors such as changing regulations, shifts in organizational culture, and broader economic conditions can play a crucial role in the implementation and sustainability of asset management processes. For example, new regulatory requirements might necessitate alterations in how assets are tracked or reported, whereas organizational culture could influence employee buy-in for new processes. Economic conditions, can also affect

available resourcesand. alter the landscape for potential improvement. Although these external elements were not explored in depth in this study, their potential influence could be underestimated.

VI. Conclusion

A. Summary of Findings

This study identified critical inefficiencies in the As-Is asset management process within ERP systems, particularly in manual asset transfers, delayed approvals, and inconsistent data synchronization in a leading Sri Lankan financial institution. A mixed-method approach integrating qualitative and quantitative analyses was employed, with data collected from 51 branches and analyzed using statistical tools such as Power BI to assess process inefficiencies, error rates, and stakeholder feedback.

The reengineered system introduces bulk asset transfers, automated approval workflows, and real-time data synchronization, delivering measurable improvements over the As-Is process. Specifically, the system achieved a 100% reduction in operational costs per user ($7,500 annual savings), an 80% reduction in compliance incidents, a 67% reduction in asset transaction errors, and a 20% improvement in asset utilization. These quantitative results demonstrate a significant enhancement in process efficiency, reduced manual effort, improved compliance, and optimized resource management.

Furthermore, the reengineering process ensures greater data accuracy, minimizes human intervention, and enhances system transparency. This transformation results in a more efficient, scalable, and user-friendly ERP asset management system, providing both academic and practical contributions by offering a validated framework for workflow optimization, process automation, and effective decision-making across large-scale enterprise environments.

B. Implications for Practice

This study highlights the crucial role of Business Process Reengineering (BPR) in optimizing enterprise resource planning (ERP) systems. The findings demonstrated that automation, real-time data synchronization, and structured workflows can significantly improve operational efficiency, regulatory compliance, and risk management. Furthermore, implementing role-based access controls enhances data security and governance,thereby reducing the likelihood of data breaches or unauthorized asset transactions.

Organizations can leverage these insights to reengineer similar processes across various business functions, including procurement, inventory control, and financial operations. This study underscores the necessity for organizations to transition from manual, labor-intensive processes to automated, data-driven systems to enhance productivity and decision-making capabilities in rapidly evolving digital environments.

C. Recommendations

In today's fast-paced business environment, effective training and change management are critical for overcoming resistance when organizations implement new systems. Comprehensive training programs designed for all users and stakeholders create a supportive learning in environment which individuals feel empowered to adopt new technology. By emphasizing the importance of training, organizations can build a culture of acceptance towards change.In Addition, a clear communication strategy that outlines the benefits of the new system is essential. When stakeholders understand how a new system can enhance their workflows, increase efficiency, and contribute to thier overall business goals, they are more likely to embrace these changes. This dual approach—robust training coupled with transparent communication—ensures that users not only comprehend the operational aspects of the new system but also feel motivated to utilize it effectively, thereby reducing disruptions during the transition period.

Instead of pursuing a full-scale rollout of new systems, organizations may find it advantageous to adopt incremental implementation strategies. This approach allows for the initial deployment of the system in the pilot phase, enabling the organization to identify potential issues in a controlled environment. By starting small, organizations can gather feedback from users, observe the system's performance, and make the necessary adjustments before expanding the implementation. This pilot phase serves as a testing ground that helps organizations fine-tune functionalities and address any unforeseen challenges, thus minimizing the risks associated with a larger rollout. Furthermore, incremental implementation fosters a sense of ownership among users, because their input during the pilot phase can lead to enhancements that cater specifically to their needs. This strategy not only facilitates smoother transitions but also enhances overall user confidence in the new system, setting the stage for a more successful organization-wide deployment.

Once the new system was implemented, continuous monitoring and evaluation were crucial to ensure that it meets the established efficiency and compliance targets. Organizations might set up regular check-ins and performance assessments to track a system's effectiveness over time. These evaluations could provide insights into how well the system functions and identify areas that require further improvement. As organizations evolve and grow, their needs may shift, necessitating ongoing adjustments to the system to align with new operational goals. By establishing a culture of continuous improvement, organizations can proactively address any shortcomings in the system and capitalize on opportunities for optimization. Furthermore, future research could explore broader applications of Business Process Reengineering (BPR) principles beyond asset management, extending into areas such as procurement and inventory control. Such studies could provide valuable insights into the transformative potential of BPR in various business functions, thereby enhancing overall organizational agility. Additionally, investigating the role of emerging technologies, such as artificial intelligence and machine learning, can open new avenues for enhancing asset management and ERP systems, driving efficiency, and improving decision-making capabilities. In conclusion, embracing a holistic approach to training, incremental implementation, and continuous evaluation can significantly mitigate resistance to change and foster a culture of adaptability in organizations.

Acknowledgment

In the preparation of this research manuscript, the only AI tool used was ChatGPT (developed by OpenAI). The

tool was employed exclusively for enhancing language clarity, improving grammar, and refining sentence structure throughout the manuscript.

To evaluate the validity and accuracy of the language outputs generated by ChatGPT, all content revised using the tool was thoroughly cross-verified by the university's English Language Unit. Additionally, the paid version of Paperpal Preflight tool was used as a supplementary resource to assess the grammatical correctness and coherence of the language improvements.

We explicitly confirm that no part of the research design, data collection, analysis, results, or interpretation was generated or influenced by ChatGPT or any other AI tools. The sole purpose of using ChatGPT was to enhance the manuscript's linguistic quality, and it was not used to create or manipulate any scientific content, findings, or supporting materials.

Furthermore, we affirm that,

- The content presented is accurate and valid, reflecting the genuine findings of the study.
- There are no concerns regarding plagiarism, and all writing is original unless explicitly cited.
- All relevant sources have been cited appropriately.
- Every statement regarding hypotheses, results, interpretations, conclusions, limitations, and implications represents the authors' own intellectual contributions and perspectives.

REFERENCES

[1] T. Prasad, "Enhancing the asset posting process in cloud enterprise resource planning application," *Journal of Economics & Management Research*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:272798355

[2] I. Lukyanova, A. Haddud, and A. Khare, "Types of erp systems and their impacts on the supply chains in the humanitarian and private sectors," *Sustainability*, vol. 14, no. 20, 2022. [Online]. Available: https://www.mdpi.com/2071-1050/14/20/13054

[3] R. Bochnia, D. Richter, and J. Anke, "Self-sovereign identity for organizations: Requirements for enterprise software," *IEEE Access*, vol. 12, pp. 7637–7660, 2024.

[4] S. AlMuhayfith and H. Shaiti, "The impact of enterprise resource planning on business performance: With the discussion on its relationship with open innovation," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 6, no. 3, 2020. [Online]. Available: https://www.mdpi.com/2199-8531/6/3/87

[5] S. Tongsuksai, S. Mathrani, and K. Weerasinghe, "Influential characteristics and benefits of cloud erp adoption in new zealand smes: A vendors' perspective," *IEEE Access*, vol. 11, pp. 23 956–23 979, 2023.

[6] L. Bartlett, M. A. Kabir, and J. Han, "A review on business process management system design: The role of virtualization and work design," *IEEE Access*, vol. 11, pp. 116 786–116 819, 2023.

[7] I. Corporation, "Business process reengineering," 2025, accessed: 2025-03-14. [Online]. Available: https://www.ibm.com/think/topics/business-process-reengineering

[8] G. H. Watson, *Business systems engineering: managing breakthrough changes for productivity and profit*. John Wiley & Sons, 1995.

[9] M. Al-Mashari and M. Zairi, "Revisiting bpr: a holistic review of practice and development," *Business process management journal*, vol. 6, no. 1, pp. 10–42, 2000.

[10] T. H. Davenport, "Putting the enterprise into the enterprise system," *Harvard Business Review*, vol. 76, no. 4, pp. 121–131, 1998, accessed: 2025-03-14. [Online]. Available: https://hbr.org/1998/07/putting-the-enterprise-into-the-enterprise-system

[11] A. Fetais, G. M. Abdella, K. N. Al-Khalifa, and A. M. Hamouda, "Business process re-engineering: A literature review-based analysis of implementation measures," *Information*, vol. 13, no. 4, 2022. [Online]. Available: https://www.mdpi.com/2078-2489/13/4/185

[12] A. Usman, "Erp systems and their impact on supply chain management," *Uncertain Supply Chain Management*, pp. 1–2, 2024.

[13] AntMyERP, "Customized eam software: Benefits of implementation," 2025, accessed: 2025-03-14. [Online]. Available: https://antmyerp.com/en/blogs/customized-eam-software-benefits-of-implementation/

[14] H. Jung and B. Kim, "Identifying research topics and trends in asset management for sustainable use: A topic modeling approach," *Sustainability*, vol. 13, no. 9, p. 4792, 2021.

[15] O. Corporation, *Oracle Financials Cloud Documentation*, 2025, accessed: 2025-03-14. [Online]. Available: https://docs.oracle.com/en/cloud/saas/financials/25a/faalm/index.html

[16] L.-E.-L. Barna, B.-S. Ionescu, and L. Ionescu-Feleaga, "The relationship between the implementation of erp systems and the financial and non-financial reporting of organizations," *Sustainability*, vol. 13, no. 21, 2021. [Online]. Available: https://www.mdpi.com/2071-1050/13/21/11566

[17] R. Kashyap, "Security framework for enterprise resource planning," in *Metrics and Models for Evaluating the Quality and Effectiveness of ERP Software*. IGI Global, 2020, pp. 84–118.

[18] Netcall, "As-is vs. to-be: Where do you start?" 2025, accessed: 2025-03-14. [Online]. Available: https://www.netcall.com/blog/as-is-vs-to-be-where-do-you-start/

[19] C. Dechsupa, W. Vatanawood, and A. Thongtak, "Stepwise verification for the bpmn with timed and stochastic process using a colored generalized stochastic petri net," *IEEE Access*, vol. 10, pp. 42 983–43 002, 2022.

[20] F. Huang, F. Ni, J. Liu, F. Yang, and J. Zhu, "A colored petri net executable modeling approach for a data flow well-structured bpmn process model," *IEEE Access*, vol. 10, pp. 86 696–86 709, 2022.

[21] A. Tarhan, O. Turetken, and H. A. Reijers, "Business process maturity models: A systematic literature review," *Information and Software Technology*, vol. 75, pp. 122–134, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584916300015

[22] R. Shah and P. T. Ward, "Defining and developing measures of lean production," *Journal of operations management*, vol. 25, no. 4, pp. 785–805, 2007.

[23] D. Kolberg and D. Zühlke, "Lean automation enabled by industry 4.0 technologies," *IFAC-PapersOnLine*, vol. 48, no. 3, pp. 1870–1875, 2015.

[24] M. L. Duc, L. Hlavaty, P. Bilik, and R. Martinek, "Design and implement low-cost industry 4.0 system using hybrid six sigma methodology for cnc manufacturing process," *IEEE Access*, vol. 11, pp. 127 176–127 201, 2023.

[25] Simplilearn, "What is six sigma? a complete overview," 2025, accessed: 2025-03-14. [Online]. Available: https://www.simplilearn.com/what-is-six-sigma-a-complete-overview-article

[26] U. Author, "Kaizen," 2025, accessed: 2025-03-14. [Online]. Available: https://www.academia.edu/16740573/Kaizen

[27] N. Koussani, A. Leshob, P. Hadaya, I. Benzarti, and H. Mili, "Towards an agile approach for business process reengineering," in *2022 IEEE International Conference on e-Business Engineering (ICEBE)*, 2022, pp. 130–135.

[28] J. S. Oakland, *Total quality management and operational excellence: text with cases*. Routledge, 2014.

[29] M. R. Spiegel and L. J. Stephens, *Schaum's Outline of Statistics*, 6th ed. New York: McGraw-Hill Education, 2018. [Online]. Available: https://www.accessengineeringlibrary.com/content/book/9781260011463

[30] O. Corporation, "Autonomous transaction processing," 2025, accessed: 2025-03-14. [Online]. Available: https://www.oracle.com/autonomous-database/autonomous-transaction-processing/

[31] I. Zafar, F. Azam, M. W. Anwar, B. Maqbool, W. H. Butt, and A. Nazir, "A novel framework to automatically generate executable web services from bpmn models," *IEEE Access*, vol. 7, pp. 93 653–93 677, 2019.

[32] Y. I. Rahmawati and A. Herawati, "The influence of good corporate governance and asset management on financial performance with erp (enterprise resource planning) as a mediating variable in the construction subsector in indonesia in 2023," *Journal of Accounting and Finance Management*, vol. 6, no. 1, pp. 201–212, 2025.

[33] C. Zeng, "Enterprise financial management in modern environment," in *2022 International Conference on Economics, Smart Finance and Contemporary Trade (ESFCT 2022)*. Atlantis Press, 2022, pp. 1116–1122.

[34] B. E. Abikoye, S. C. Umeorah, A. O. Adelaja, O. Ayodele, and Y. M. Ogunsuji, "Regulatory compliance and efficiency in financial technologies: Challenges and innovations," *World Journal of Advanced Research and Reviews*, vol. 23, no. 1, pp. 1830–1844, 2024.

[35] N. Farhadighalati, L. A. Estrada-Jimenez, S. Nikghadam-Hojjati, and J. Barata, "A systematic review of access control models: Background, existing research, and challenges," *IEEE Access*, vol. 13, pp. 17 777–17 806, 2025.

[36] J. C. Bauer, E. John, C. L. Wood, D. Plass, and D. Richardson, "Data entry automation improves cost, quality, performance, and job satisfaction in a hospital nursing unit," *JONA: The Journal of Nursing Administration*, vol. 50, no. 1, pp. 34–39, 2020.

[37] R. Pushpakumar, K. Sanjaya, S. Rathika, A. H. Alawadi, K. Makhzuna, S. Venkatesh, and B. Rajalakshmi, "Human-computer interaction: Enhancing user experience in interactive systems," in *E3S Web of Conferences*, vol. 399. EDP Sciences, 2023, p. 04037.

[38] I. Pastierik, "Deploying oracle machine learning automl models for oracle apex analytics," in *2024 IEEE 17th International Scientific Conference on Informatics (Informatics)*, 2024, pp. 499–506.

[39] O. Corporation, *Oracle E-Business Suite Documentation*, 2025, accessed: 2025-03-14. [Online]. Available: https://docs.oracle.com/cd/E26401_01/doc.122/f21671/T673609T683724.htm

[40] O. L. Experts, "Oracle erp cloud licensing costs," 2025, accessed: 2025-03-14. [Online]. Available: https://oraclelicensingexperts.com/oracle-erp-cloud-licensing-costs/

Stochastic Policies, Deterministic Minds: A Calibrated Evaluation Protocol and Diagnostics for Deep Reinforcement Learning

Sooyoung Jang[1], Seungho Yang[2], Changbeom Choi[3]*
Department of Computer Engineering, Hanbat National University, Daejeon, 34158, Republic of Korea[1,3]
Department of Urban Engineering, Hanbat National University, Daejeon, 34158, Republic of Korea[2]

***Abstract*—Deep reinforcement learning (DRL) typically involves training agents with stochastic exploration policies while evaluating them deterministically. This discrepancy between stochastic training and deterministic evaluation introduces a potential objective mismatch, raising questions about the validity of current evaluation practices. Our study involved training 40 Proximal Policy Optimization agents across eight Atari environments and examined eleven evaluation policies ranging from deterministic to high-entropy strategies. We analyzed mean episode rewards and their coefficient of variation while assessing one-step temporal-difference errors related to low-confidence actions for value-function calibration. Our findings indicate that the optimal evaluation policy is highly dependent on the environment. Deterministic evaluation performed best in three games, while low-to-moderate-entropy policies yielded higher returns in five, with a significant improvement of over 57% in Breakout. However, increased policy entropy generally degraded stability—evidenced by a rise in the coefficient of variation in Pong from 0.00 to 2.90. Additionally, low-confidence actions often revealed an over-optimistic value function, exemplified by negative TD errors, including -10.67 in KungFuMaster. We recommend treating evaluation-time entropy as a tunable hyperparameter, starting with deterministic or low-temperature softmax settings to optimize both return and stability on held-out seeds. These insights provide actionable strategies for practitioners aiming to enhance their DRL-based agents.**

***Keywords*—Deep reinforcement learning; policy evaluation; stochastic policy; temporal difference error; Atari; PPO**

I. Introduction

Deep reinforcement learning has emerged as a powerful paradigm for solving complex sequential decision-making problems, with notable successes in domains ranging from game playing to robotics [1], [2]. Among the most successful families of algorithms are policy gradient methods, which directly optimize the parameters θ of an agent's policy, π_θ. Algorithms like Proximal Policy Optimization (PPO) [3] learn a stochastic policy, which maps states to a probability distribution over actions, $\pi_\theta(a|s)$. For discrete action spaces, as found in the Atari Learning Environment, this policy is typically represented by a categorical distribution, often parameterized by the softmax output of a neural network.

The stochastic nature of the policy is not an incidental feature; it is fundamental to the learning process. During training, sampling actions from the policy distribution enables exploration, allowing the agent to discover novel state-action pathways that may lead to higher rewards. This exploration is essential for escaping local optima and preventing premature convergence to a suboptimal policy. Indeed, the policy gradient theorem [4], the theoretical foundation of these methods, relies on this stochasticity to estimate the gradient of the expected return.

In stark contrast to the training phase, the common practice during evaluation or deployment is to render the policy deterministic by selecting the action with the highest probability for any given state (i.e., taking the argmax of the policy's output distribution). This deterministic approach is predicated on the principle of exploitation. Once the agent has been sufficiently trained, the optimal strategy is to consistently choose the action it believes to be the best. This creates a fundamental duality: agents are trained as stochastic decision-makers but are often deployed as deterministic ones. This practice, while widespread, rests on an assumption that has received surprisingly little systematic scrutiny. Earlier empirical studies have examined evaluation discrepancies for value-based agents, but only recently has the gap for policy-gradient methods been quantified. In study [5], the authors demonstrate that apparent gains can disappear when stochastic and deterministic evaluations are compared under matched conditions; however, their study does not isolate the causal role of value-function mis-calibration. We extend this thread by providing the first large-scale analysis that links evaluation stochasticity, TD-error structure, and overfitting across eight Atari domains.

This dichotomy between training and evaluation methodologies raises a critical question: is the deterministic policy derived via an argmax operation the most effective implementation of the knowledge encoded within the trained stochastic policy? While intuitively appealing, this assumption is not guaranteed to hold. The optimization objective of PPO maximizes the expected return under the stochastic policy, π_θ, not necessarily the return of its deterministic counterpart. This creates a potential objective mismatch, where the criteria for success during training (effective stochastic exploration and exploitation) differ from the criteria during evaluation (pure deterministic exploitation). There are well-established scenarios where a stochastic policy is not just a tool for exploration but is fundamentally optimal for the task itself. In game-theoretic or multi-agent settings, a deterministic policy can be predictable and easily exploited by an adversary. A classic example is Rock-Paper-Scissors, where the Nash equilibrium corresponds to a uniform stochastic policy [4]. Similarly, in Partially Observable Markov Decision Processes (POMDPs),

*Corresponding author.

where different underlying states may appear identical to the agent (a phenomenon known as perceptual aliasing), a stochastic policy can be crucial for breaking symmetry and avoiding getting stuck. Although Atari games are typically single-agent environments, they are not fully observable, as the agent only sees the current screen's pixels and must infer dynamic information like velocity and trajectory. This partial observability could favor a stochastic approach. The degree of a policy's randomness can be quantified by its entropy. Higher entropy corresponds to a more uniform and exploratory distribution, while lower entropy indicates a more deterministic and exploitative one. During training, policy entropy is often explicitly encouraged via an entropy bonus in the loss function to promote exploration and smooth the optimization landscape. However, its role and optimal level during the evaluation phase remain largely unexamined.

This paper aims to bridge this research gap through a large-scale, systematic empirical investigation. We trained 40 agents using the PPO algorithm across eight distinct Atari environments. These trained agents were then evaluated using a suite of eleven different evaluation policies, ranging from fully deterministic to highly stochastic, allowing us to precisely control and measure the impact of policy entropy on performance. This comprehensive experimental framework allows us to address the following research questions:

- Performance and Stability: How do evaluation policies with varying degrees of stochasticity, as measured by entropy, affect agent performance (mean reward) and stability (reward variance) across different Atari environments?
- Action Pruning: Can a simple heuristic—pruning the least likely actions from the policy's distribution, a form of manual entropy reduction—consistently improve performance?
- Underlying Mechanisms: How do stochastic policies reveal weaknesses in the learned value function? Specifically, can we use the Temporal Difference (TD) error—conceptually, the difference between the predicted value of a state and a more accurate value estimated after taking an action— on low-confidence actions to diagnose inaccurate value predictions and explain performance degradation?

Novelty and contribution. In contrast to prior work that either 1) proposes entropy-regularized objectives to train intrinsically stochastic policies [6], [7] or 2) benchmarks evaluation heuristics without examining the critic, our study treats evaluation stochasticity as a post-training intervention and interrogates its causal impact on performance through a TD-error lens. This framing enables a principled diagnosis of value-function overfitting in cases where stochasticity harms performance and yields actionable guidelines for selecting evaluation protocols that account for this risk.

The remainder of this paper is structured to answer these questions. Section II details the related works, situating our investigation within the context of policy representation and overfitting in RL. Section III describes the methodologies we utilized for the study. Section IV presents a detailed analysis of the experimental results, addressing each research question in turn. Finally, Section V discusses the broader implications of our findings for RL practitioners and outlines directions for future research, concluding with a summary in Section VI.

II. Related Work

Our investigation is situated at the intersection of several key areas in reinforcement learning: policy representation, overfitting and generalization, action space modification, and the role of stochasticity.

A. Policy Representation in Reinforcement Learning

RL algorithms can be broadly categorized by how they represent the agent's policy.

A core distinction in RL is between on-policy and off-policy learning. On-policy algorithms, such as PPO [3], update the policy using only the data collected from the most recent version of that same policy. In contrast, off-policy algorithms, like DDPG [8], can learn from data generated by a different policy, often by using a replay buffer of past experiences. Another key distinction is between stochastic policies, which map states to a probability distribution over actions, and deterministic policies, which map states to a single action. These two distinctions are not synonymous.

1) Stochastic policy methods: Modern on-policy RL was significantly advanced by Trust Region Policy Optimization (TRPO) [9], which introduced a constraint on the Kullback-Leibler (KL) divergence between successive policies to guarantee monotonic improvement and stabilize training. PPO [3] emerged as a more practical and scalable successor, replacing the complex second-order optimization of TRPO with a more straightforward clipped surrogate objective that achieves similar stability. Its robustness and ease of implementation have made PPO a ubiquitous algorithm in the field, with recent advances continuing to refine its optimization process and theoretical underpinnings [10], [11]. Another influential on-policy method, Asynchronous Advantage Actor-Critic (A3C) [12], demonstrated that using multiple parallel actors to interact with the environment could decorrelate the training data and stabilize learning without a replay buffer. These methods all directly learn a stochastic policy, where exploration is inherent to the policy representation itself. While these examples are on-policy, off-policy algorithms that learn stochastic policies, like Soft Actor-Critic (SAC) [7], also exist.

2) Deterministic policy methods: In contrast, another line of research focuses on learning deterministic policies. The Deterministic Policy Gradient (DPG) theorem [13] showed that for deterministic policies, the policy gradient can be computed more efficiently as it does not require integration over the action space. This theoretical advantage motivated the development of algorithms like Deep DPG (DDPG) [8], which combines DPG with deep function approximators for continuous control tasks. These methods are typically off-policy and separate exploration from the policy itself, typically by adding noise to the actions during training only.

This fundamental split in RL approaches—between learning a stochastic policy with built-in exploration and learning a deterministic policy with external exploration—sets the stage for our central research question.

3) Value-based methods and evaluation: The pioneering Deep Q-Network (DQN) [1] learns a value function, Q(s,a), and derives a policy by acting greedily with respect to it. Early work on DQN evaluated agents using an ϵ-greedy policy with a small ϵ (e.g., 0.05) to prevent deterministic cycling in the Atari environments. However, subsequent common practice shifted towards fully greedy evaluation ($\epsilon = 0$) to assess the agent's maximum performance capacity. This historical trend shows the field's implicit convergence on deterministic evaluation as the standard, an assumption our work aims to scrutinize. Complementary evidence comes from [5], who revisit the Atari evaluation methodology and advocate measuring on-policy returns under test-time stochasticity. Our experimental design follows these recommendations but augments them with a TD-error-based causal probe, thereby clarifying why mismatched evaluation protocols degrade performance.

B. Overfitting in Deep Reinforcement Learning

The problem of generalization, a fundamental challenge in deep reinforcement learning (DRL) [14], provides a critical lens through which to view our results. High-capacity models like deep neural networks are prone to memorizing the training data rather than learning generalizable features, a phenomenon that is particularly acute in RL. An agent that overfits to the limited set of trajectories collected through its own on-policy interactions may perform well on states and actions it has seen frequently but fail catastrophically in novel situations [15]. For example, a study [16] demonstrated that overfitting is a primary cause of performance collapse when an agent is transferred from a limited set of training levels to unseen ones, highlighting the need for explicit regularization techniques to promote generalization.

Recent work has provided a more specific vocabulary for diagnosing these failures. The issue is not merely one of standard supervised learning overfitting, but involves instabilities unique to the RL process. A key failure mode is the chain effect of churn [17]. This phenomenon, identified in [17], describes how the value function's predictions for states not in the batch can change uncontrollably after each batch update. This creates a vicious cycle where churn in value estimation and policy improvement compound each other, progressively biasing the learning dynamics. This churn is symptomatic of a deeper issue termed "representation collapse," where the agent's neural networks lose plasticity and feature rank throughout training, rendering them incapable of adapting to new observations or fitting new targets. This collapse is exacerbated by the non-stationarity inherent in RL, where the agent's improving policy continually shifts the data distribution.

This perspective aligns with the concept of specification overfitting, where a system focuses excessively on optimizing a specific, narrow metric—in this case, the TD error on its on-policy data distribution—to the detriment of broader, high-level requirements, such as robustness and generalization. In [18], the authors provided a crucial diagnostic insight, identifying statistical overfitting in the temporal-difference (TD) error as a primary bottleneck. Their key idea is that as an agent trains, its value function becomes highly accurate on its on-policy data distribution but remains inaccurate on out-of-distribution transitions. They propose that this can be diagnosed by measuring the TD error on a held-out validation set of transitions, where an extensive validation TD error indicates poor generalization. This theoretical framework provides a powerful explanation for our empirical findings. The low-confidence actions sampled by our stochastic evaluation policies serve as a naturally occurring validation set. The TD errors on these actions thus act as a direct probe into the value function's generalization capabilities, allowing us to empirically test and validate the overfitting hypothesis proposed in the recent literature.

C. Action Elimination and Pruning

Our experiments with drop policies are related to a line of research on action elimination. An Action Elimination Network (AEN) learns to predict and mask out suboptimal or irrelevant actions in environments with very large discrete action spaces [19]. By reducing the effective size of the action space at each step, their method significantly speeds up learning and improves robustness in complex domains like text-based games. Our drop policies can be seen as a simple, non-learned heuristic inspired by this principle. Instead of learning which actions to eliminate, we manually prune the actions with the lowest logits, allowing us to test the hypothesis that performance can be improved simply by preventing the agent from taking actions it is already least confident about.

D. The Dual Role of Stochasticity: Exploration, Optimality, and Causality

The stochastic nature of the policy in algorithms like PPO is typically framed as a mechanism for exploration, which is essential for discovering novel, high-reward trajectories and preventing premature convergence. However, this view is incomplete. A growing body of research highlights that policy stochasticity plays a more fundamental role in optimization and even optimality itself. In some contexts, a stochastic policy is not merely a means to an end but is the optimal solution. This is well-established in game-theoretic settings [4]. More broadly, the framework of maximum entropy RL formalizes the benefit of randomness by augmenting the standard reward objective with an entropy term [6]. Algorithms like SAC [7] explicitly optimize this objective, learning policies that are not only effective but also as random as possible. The resulting policies can capture diverse strategies for accomplishing a task, thereby improving robustness.

Furthermore, the stochasticity inherent in policy gradient methods is a key mechanism for escaping local optima in the complex, non-convex optimization landscapes of DRL [20]. A purely deterministic policy can easily get trapped in a simple but suboptimal strategy. In contrast, the noise introduced by sampling from a policy distribution can provide the necessary perturbation to jolt the agent out of such traps and discover more sophisticated, higher-reward solutions. This provides a compelling theoretical lens through which to interpret our empirical results in environments like Breakout.

Finally, recent work on the causal foundations of RL provides a formal way to disentangle the sources of randomness in an agent's experience [21]. This framework decomposes the total return of a trajectory into components attributable to the agent's actions and components attributable to the environment's inherent stochasticity. This decomposition is achieved by interpreting the advantage function as the causal

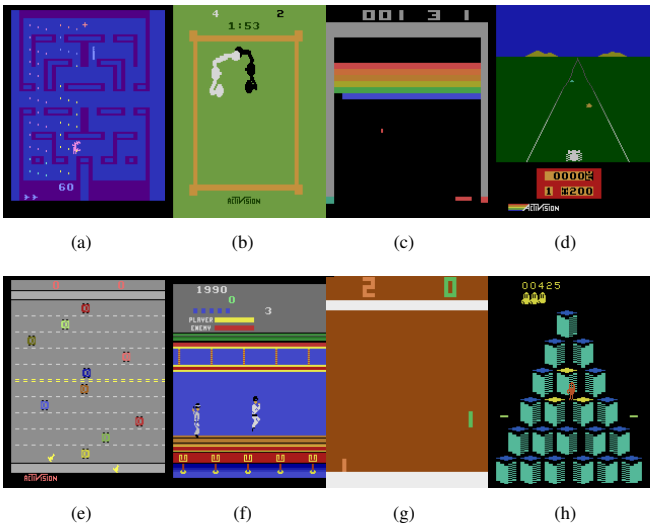


Fig. 1. The eight Atari 2600 environments utilized for the study: (a) Alien, (b) Boxing, (c) Breakout, (d) Enduro, (e) Freeway, (f) KungFuMaster (KFM), (g) Pong, and (h) Qbert.

effect of an action on the return [21]. This distinction clarifies that our investigation focuses on the consequences of the agent's policy stochasticity rather than randomness in the environment's dynamics.

III. Experimental Framework

To systematically investigate the effects of evaluation policy stochasticity, we designed a comprehensive experimental setup involving multiple agents, environments, and evaluation strategies.

A. Environments and Agent Training

Experiments were performed on eight Atari 2600 games from the Arcade Learning Environment (ALE)—Alien, Boxing, Breakout, Enduro, Freeway, KungFuMaster (KFM), Pong, and Qbert (Fig. 1). These titles span a broad spectrum of interaction modalities, reward densities, and action-space cardinalities, thereby providing a representative benchmark suite. The characteristics of each environment, particularly the size of the discrete action space and the corresponding entropy of a random policy, are crucial for interpreting the results and are summarized in Table I.

For each game, five independent PPO agents were trained, yielding a total of 40 trained policies and thereby mitigating the variance induced by random seeds. Training lasted for 1×10^7 environment steps per agent and used the canonical convolutional network for Atari inputs introduced by study [1]. Table II lists the set of PPO hyperparameters that were held constant across all environments.

B. Evaluation Policies

For each of the 40 trained agents, we conducted evaluations using 11 distinct action selection policies. These policies were designed to span a wide spectrum of stochasticity, from fully deterministic to highly random. Each agent-policy combination was evaluated for 100 independent episodes to gather statistically robust performance data. The definitions of the evaluation policies are provided in Table III. Table III provides a concise definition for each of the 11 evaluation policies used to assess the trained agents. logits refers to the raw, pre-softmax output of the policy network.

Justification for linear and drop policies:

1) Linear: The linear policy was designed as a high-entropy alternative to a purely random policy. By preserving the ordinal ranking of the network's logits while drastically flattening the probability distribution, it serves as a hard test case for how the agent performs when it must consider its less-preferred actions more seriously.

2) Drop: The softmax_drop_k and linear_drop_k policies are simple, non-learned heuristics inspired by action-pruning techniques. They allow us to test the hypothesis that manually preventing the agent from taking actions it is already least confident about can function as a form of targeted entropy reduction and improve performance, particularly on otherwise highly random policies.

C. Analysis Metrics

To evaluate performance and diagnose mechanisms, we record the following metrics during evaluation.

1) Performance metrics: The primary performance metric is the *Mean Episode Reward*, averaged over 100 episodes per agent–policy pair. Stability is quantified by the *Coefficient of Variation* (CV), defined as the standard deviation of episode returns divided by their mean. A lower CV indicates more consistent performance relative to the average score.

2) Stochasticity metric: The degree of randomness of the evaluation policy is measured by the average policy entropy,

$$H(q_T) \;=\; \mathbb{E}_t\left[-\sum_{a\in A} q_T(a \mid s_t)\,\log q_T(a \mid s_t)\right], \qquad (1)$$

where $q_T(\cdot \mid s)$ is the evaluation policy (e.g., `softmax_T`) and the expectation is over timesteps.

3) Diagnostic metrics: To investigate the underlying mechanisms of performance differences, we defined two key diagnostic metrics:

a) Low-Confidence Action Ratio (LCAR): Given an evaluation policy $q_T(a \mid s)$ and threshold $\tau \in (0, 1]$, define

$$\mathrm{LCAR}_\tau(s) \;=\; \sum_{a\in A} q_T(a \mid s)\,\mathbf{1}\Big\{\, q_T(a \mid s) < \tfrac{\tau}{|A|}\Big\}. \qquad (2)$$

This quantity equals the total probability mass that $q_T(\cdot \mid s)$ assigns to the *low-confidence* set $\{a : q_T(a \mid s) < \tau/|A|\}$. We report **LCAR-50** ($\tau = 0.5$) and **LCAR-10** ($\tau = 0.1$) as the time-averaged frequencies of such events. *Note on softmax_10.0.* As $T \to \infty$, $q_T(\cdot \mid s)$ approaches the uniform distribution, implying $\mathrm{LCAR}_\tau(s) \to 0$ because $1/|A| > \tau/|A|$

TABLE I. ACTION SPACE SIZE AND ENTROPY OF A RANDOM POLICY IN ATARI ENVIRONMENTS

	Alien	Boxing	Breakout	Enduro	Freeway	KFM	Pong	Qbert
Action Space Size (N)	18	18	4	9	3	14	6	6
Entropy (nats)	2.890	2.890	1.386	2.197	1.099	2.639	1.792	1.792

TABLE II. PPO HYPERPARAMETER CONFIGURATION USED FOR THE ATARI EXPERIMENTS

Hyper-parameter	**Value**
Frame stack	4
Number of parallel environments (n_{envs})	8
Rollout length per environment (n_{steps})	128
SGD epochs per update (n_{epochs})	4
Mini-batch size	256
Total timesteps	1×10^7
Learning-rate schedule	linear, $2.5 \times 10^{-4} \to 0$
Value-function loss coefficient (c_v)	0.5
Entropy coefficient (c_H)	0.01

TABLE III. EVALUATION POLICIES AND THEIR DESCRIPTIONS

Policy	**Description**
deterministic	Selects the action with the highest logit value: $a = \arg\max(\text{logits})$.
softmax_T	Samples an action from the distribution given by $\text{Softmax}(\text{logits}/T)$. Tested for temperatures $T \in \{0.1, 0.5, 1.0, 2.0, 10.0\}$.
linear	Applies an affine shift to make all logits non-negative ($w = v - \min(v)$), then normalizes these values linearly to form a probability distribution.
softmax_drop_k	Removes the k actions with the lowest logit values, then applies a softmax function with $T = 1.0$ to the remaining logits. Tested for $k \in \{1, 2\}$.
linear_drop_k	Removes the k actions with the lowest logit values, then applies the linear normalization scheme to the remaining logits. Tested for $k \in \{1, 2\}$.

for $\tau \in \{0.5, 0.1\}$. At the finite temperature T=10, however, the policy remains only approximately uniform and retains residual structure from the logits. Consequently, a small but nonzero fraction of probability mass can still fall below the threshold. Empirically, this is visible in Qbert, where softmax_10.0 attains LCAR-50 = 5.489% (Table XI). We therefore include softmax_10.0 as a high-entropy reference policy rather than a recommended operating point.

b) Conditional TD error: The one-step TD error at timestep t is

$$\delta_t \;=\; R_{t+1} + \gamma\, V(S_{t+1}) - V(S_t), \tag{3}$$

with discount factor $\gamma = 0.99$ and V the learned value function. We report the average TD error *conditioned* on low-confidence events (e.g., timesteps contributing to LCAR-50 or LCAR-10). This conditional analysis probes the value function's accuracy for actions that the policy itself deems unlikely.

IV. RESULTS AND ANALYSIS

Our experiments yielded a rich dataset that provides clear answers to our research questions, with comprehensive statistics presented in Tables IV to XI for each environment and Table XII for policy entropies. The results, visualized in Fig. 2, reveal a split pattern: in five environments, a low- or moderate-entropy policy outperforms the deterministic baseline, while in three (Boxing, Pong, Qbert) deterministic remains best.

A. RQ1: How do Stochastic vs. Deterministic Policies Affect Performance and Stability?

1) Performance: Across the eight Atari tasks, deterministic evaluation achieves the highest mean reward in three games—Boxing (Table V), Pong (Table X), and Qbert (Table XI). In the remaining five games, a low-to-moderate-entropy policy outperforms deterministic evaluation: Alien (softmax_0.5; Table IV), Breakout (softmax_0.5; Table VI), Enduro (softmax_0.1; Table VII), Freeway (softmax_drop_1; Table VIII), and KFM (softmax_0.1; Table IX). This pattern indicates that modest stochasticity can yield higher returns in a majority of environments, while deterministic remains a strong baseline in the others. This pattern is evident in the summary bar charts in Fig. 2. In several environments, modest stochasticity yields higher returns. In Breakout (Table VI), the moderately stochastic softmax_0.5 policy surpasses the deterministic baseline, scoring 242.02 compared to 153.83 (57.3% improvement). In Freeway (Table VIII), performance differences are modest, but softmax_1.0 and drop variants slightly exceed the deterministic baseline (22.00 and 22.06 vs 21.41), indicating multiple near-optimal evaluation policies. However, in Qbert (Table XI), the deterministic policy scores 3894.00, significantly outperforming the standard stochastic softmax_1.0 policy, which scores 2404.60.

2) Stability: Stochastic evaluation generally inflates return variance and degrades stability. Across many games, higher-entropy evaluation tends to increase the coefficient of variation (CV). However, the relationship is not strictly monotonic in all environments, so we treat the stability impacts of entropy as environment-dependent(as detailed in Table XII). This instability is most pronounced in what we term 'brittle' environments, such as Pong (Table X), where a single suboptimal action can lead to immediate and irreversible negative consequences (e.g., losing a rally). The CV rises from a perfect 0.00 under the deterministic policy to 0.38 under the softmax_2.0 policy. Conversely, in 'robust' environments like Freeway (Table VIII), the agent's success is determined by a longer sequence of actions, and a single suboptimal move is less likely to be catastrophic, allowing for multiple near-optimal policies. However, stability decreases significantly with near-uniform action sampling, as seen with the softmax_10.0, which has a CV of 0.42.

To better understand the differences in both performance and stability, Fig. 3 visualizes the full reward distributions for

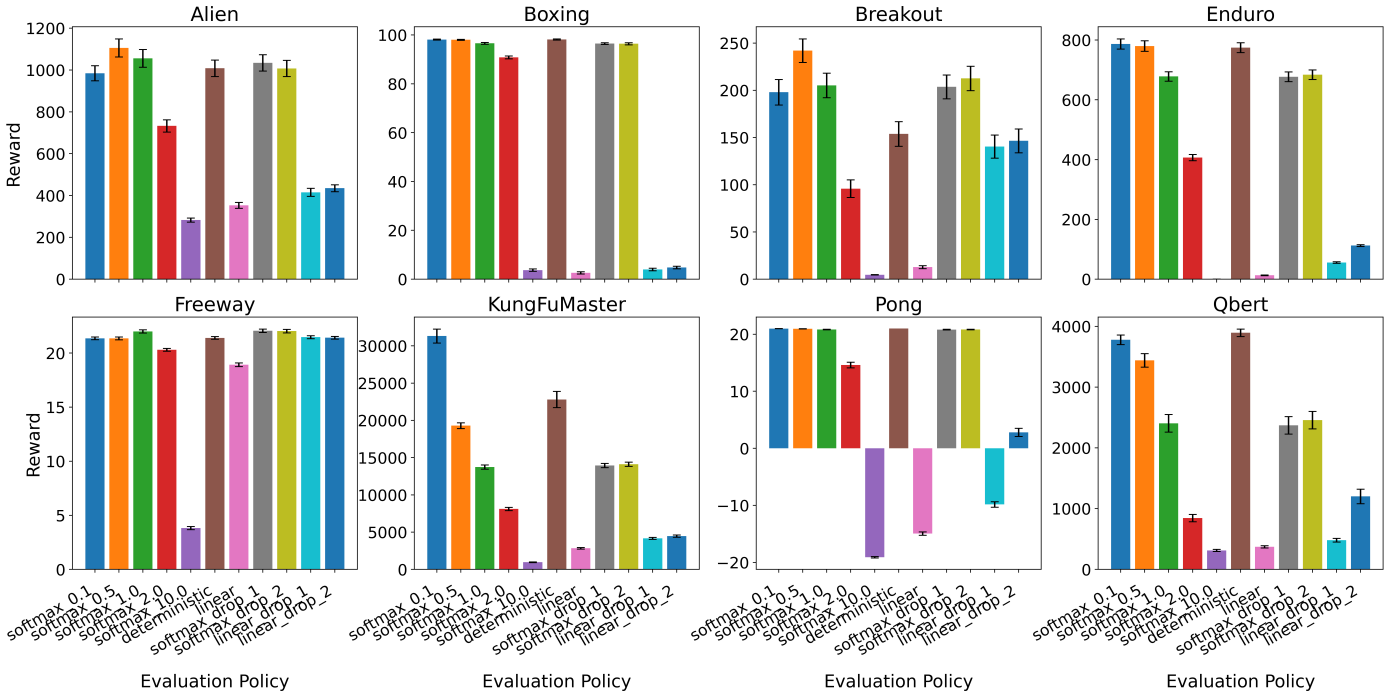


Fig. 2. Mean reward per evaluation policy across eight Atari environments. Error bars represent 95% confidence intervals.

TABLE IV. POLICY-WISE SUMMARY FOR ALIEN: REWARD, LOW CONFIDENCE ACTION RATIO, AND TD ERROR

Policy	Reward				Low Conf 50%		Low Conf 10%	
	Mean	Min	Max	CV	Ratio (%)	TD Error	Ratio (%)	TD Error
softmax_0.1	984.46	460.00	2850.00	0.41	0.489	0.765	0.100	0.518
softmax_0.5	1105.72	470.00	4400.00	0.44	3.075	0.205	0.534	-0.014
softmax_1.0	1055.94	330.00	4080.00	0.46	5.111	0.039	0.835	-0.253
softmax_2.0	732.78	280.00	2860.00	0.45	7.521	-0.497	0.896	-0.983
softmax_10.0	282.78	70.00	1450.00	0.40	2.567	-1.819	0.002	1.368
deterministic	1008.28	470.00	2650.00	0.44	0.000	0.000	0.000	0.000
linear	352.96	80.00	2090.00	0.45	5.438	-1.236	0.117	-0.796
softmax_drop_1	1034.18	270.00	2910.00	0.43	5.024	0.041	0.832	0.059
softmax_drop_2	1007.52	330.00	4370.00	0.44	5.050	0.028	0.826	-0.395
linear_drop_1	415.36	150.00	2170.00	0.53	4.697	-0.923	0.173	-1.274
linear_drop_2	435.18	90.00	1970.00	0.43	4.145	-0.910	0.176	-0.933

the deterministic policy and a representative stochastic policy (softmax_1.0). These plots reveal the underlying structure of the summary statistics. For example, in Pong, the deterministic policy's distribution is a single sharp spike at the maximum score of 21, illustrating its perfect stability (CV of 0.00). In contrast, the distribution for Breakout is bimodal; the stochastic softmax_1.0 policy is visibly more successful at avoiding the low-reward mode, explaining its higher average score. Conversely, in games like Qbert, the deterministic policy's distribution is clearly concentrated at a much higher reward level than that of the stochastic policy, reinforcing its superior performance in that environment.

B. RQ2: What is the effect of action pruning (drop) on performance?

Our analysis of the drop policies, which manually prune the least likely actions, reveals that this heuristic acts as a form of manual entropy reduction whose effectiveness is entirely dependent on the baseline policy's randomness.

- When applied to a focused, low-entropy policy like softmax_1.0, pruning has a negligible impact. The actions being pruned already have near-zero probability of being selected, so their explicit removal does not significantly alter the agent's behavior. For instance, in Alien (Table IV), the mean reward for softmax_1.0 is 1055.94, while softmax_drop_1 and softmax_drop_2 score a comparable 1034.18 and 1007.52, respectively.
- The effect is dramatically different when pruning is applied to a highly random policy such as linear, which assigns substantial probability to actions with low logits. In this context, pruning is highly effective at improving performance by forcing the agent to be more exploitative. In Qbert (Table XI), the linear policy scores a meager 371.55, whereas linear_drop_2 improves this to 1201.20, more than tripling the reward. This shows that performance is fundamentally

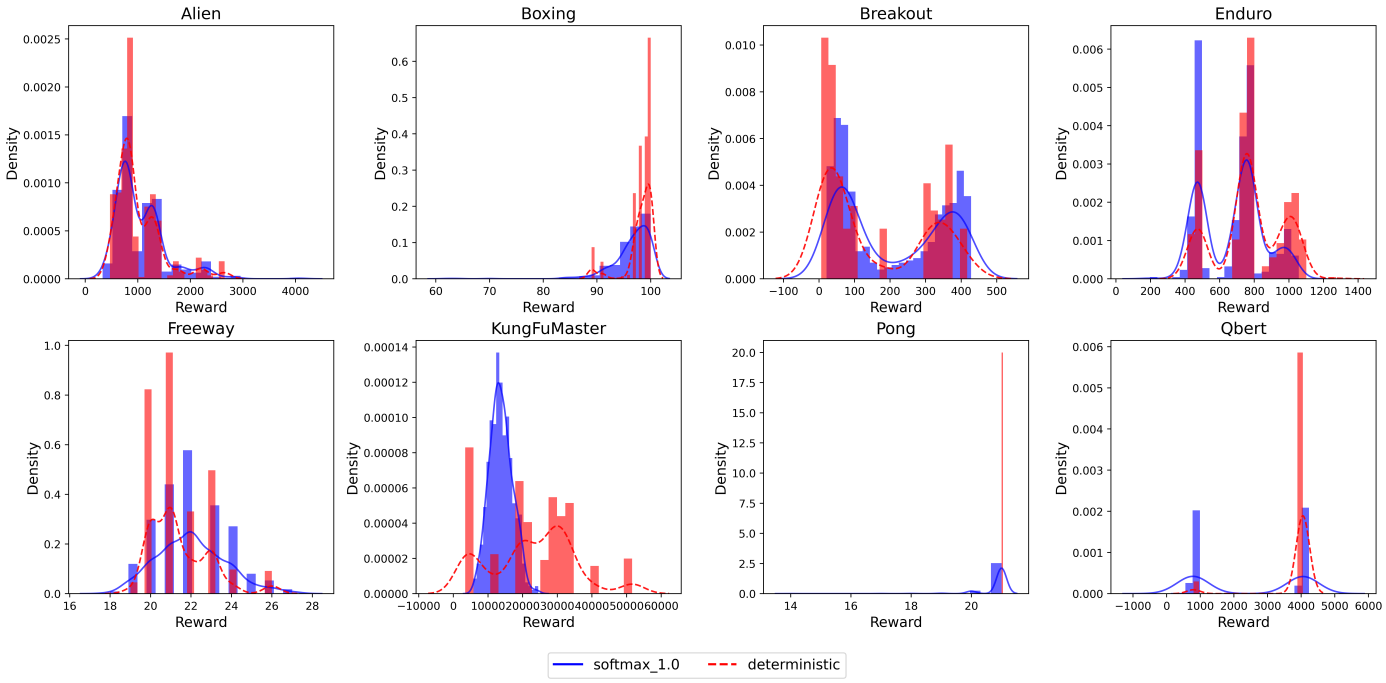


Fig. 3. Reward distribution histograms and kernel density estimates for 100 episodes across eight Atari environments. Only two representative policies (softmax policy with temperature 1.0 and deterministic policy) are shown for clarity.

TABLE V. POLICY-WISE SUMMARY FOR BOXING: REWARD, LOW CONFIDENCE ACTION RATIO, AND TD ERROR

Policy	Reward				Low Conf 50%		Low Conf 10%	
	Mean	Min	Max	CV	Ratio (%)	TD Error	Ratio (%)	TD Error
softmax_0.1	98.11	84.00	100.00	0.02	0.098	0.006	0.020	0.030
softmax_0.5	98.02	84.00	100.00	0.02	0.706	-0.011	0.146	-0.010
softmax_1.0	96.57	62.00	100.00	0.04	1.777	-0.022	0.422	-0.032
softmax_2.0	90.82	42.00	100.00	0.07	5.148	-0.078	1.065	-0.114
softmax_10.0	3.72	-32.00	22.00	1.52	4.056	-0.170	0.010	-0.328
deterministic	98.17	89.00	100.00	0.03	0.000	0.000	0.000	0.000
linear	2.59	-18.00	22.00	2.02	3.925	-0.139	0.074	-0.188
softmax_drop_1	96.48	66.00	100.00	0.04	1.769	-0.032	0.420	-0.046
softmax_drop_2	96.41	46.00	100.00	0.05	1.814	-0.035	0.433	-0.048
linear_drop_1	3.99	-20.00	25.00	1.53	4.896	-0.144	0.120	-0.163
linear_drop_2	4.80	-9.00	25.00	1.22	4.844	-0.137	0.142	-0.153

tied to avoiding low-confidence actions. However, it is crucial to note that even with this significant improvement, the linear_drop_2 policy's score of 1201.20 is still dramatically lower than the deterministic policy's score of 3894.00. This suggests that simply pruning the worst few actions is insufficient to recover high performance. The underlying issue is not just the presence of a few catastrophic actions, but the overall shape of the probability distribution. The linear normalization scheme, even after pruning, assigns a much more uniform probability to the remaining actions than the peaked distribution of a standard softmax policy, leading to continued suboptimal decision-making.

C. RQ3: Can TD Error Explain the Performance Degradation?

While the preceding results demonstrate a correlation between high evaluation entropy and poor returns in many games, our analysis of the TD error provides a causal, mechanistic explanation for this performance degradation. The data reveal that performance degradation is a direct symptom of the agent's value function being poorly generalized outside of the high-density regions of the on-policy data distribution observed during training. The argument is built in three steps.

1) Step 1: Higher entropy forces more low-confidence actions. There is a direct causal link between the stochasticity of the evaluation policy and the frequency of low-confidence actions. As policy entropy increases (Table 12), the evaluation policy q_T allocates more probability to the low-probability tail, increasing the frequency of low-confidence events (e.g., LCAR-50 of Alien rises from 0.489% at $T = 0.1$ to 7.521% at $T = 2.0$; Table IV).

2) Step 2: The value function is systematically overoptimistic for low-confidence actions. Our analysis shows that when the agent is forced to take actions it has learned to avoid, the resulting outcomes are consistently worse than what its

TABLE VI. POLICY-WISE SUMMARY FOR BREAKOUT: REWARD, LOW CONFIDENCE ACTION RATIO, AND TD ERROR

Policy	Reward				Low Conf 50%		Low Conf 10%	
	Mean	Min	Max	CV	Ratio (%)	TD Error	Ratio (%)	TD Error
softmax_0.1	198.01	4.00	426.00	0.78	1.593	-0.122	0.311	-0.200
softmax_0.5	242.02	22.00	429.00	0.59	3.454	0.033	0.524	0.029
softmax_1.0	205.22	21.00	428.00	0.72	5.153	-0.008	0.667	-0.137
softmax_2.0	95.88	7.00	421.00	1.11	6.478	-0.127	0.529	-0.356
softmax_10.0	4.71	0.00	16.00	0.65	1.876	-0.180	0.015	-0.134
deterministic	153.83	6.00	417.00	0.96	0.000	0.000	0.000	0.000
linear	12.79	0.00	373.00	1.38	1.269	-0.032	0.029	-0.056
softmax_drop_1	203.68	15.00	430.00	0.70	5.540	-0.016	2.268	-0.020
softmax_drop_2	212.63	17.00	429.00	0.69	10.709	0.000	9.711	0.005
linear_drop_1	140.39	7.00	428.00	0.99	1.032	-0.091	0.037	-0.073
linear_drop_2	146.49	6.00	417.00	0.98	0.000	0.000	0.000	0.000

TABLE VII. POLICY-WISE SUMMARY FOR ENDURO: REWARD, LOW CONFIDENCE ACTION RATIO, AND TD ERROR

Policy	Reward				Low Conf 50%		Low Conf 10%	
	Mean	Min	Max	CV	Ratio (%)	TD Error	Ratio (%)	TD Error
softmax_0.1	786.96	390.00	1341.00	0.25	0.789	0.000	0.162	0.009
softmax_0.5	779.99	377.00	1359.00	0.26	3.652	-0.017	0.586	-0.027
softmax_1.0	678.18	197.00	1057.00	0.26	5.346	-0.041	0.728	-0.073
softmax_2.0	406.85	135.00	756.00	0.29	6.203	-0.103	0.824	-0.170
softmax_10.0	0.50	0.00	14.00	3.51	2.901	-0.271	0.001	-0.316
deterministic	774.72	415.00	1272.00	0.24	0.000	0.000	0.000	0.000
linear	13.17	0.00	107.00	0.98	4.072	-0.197	0.273	-0.195
softmax_drop_1	677.04	389.00	1057.00	0.27	5.334	-0.041	0.927	-0.065
softmax_drop_2	684.12	192.00	1083.00	0.26	5.396	-0.039	1.266	-0.051
linear_drop_1	55.74	0.00	140.00	0.54	2.297	-0.187	0.166	-0.224
linear_drop_2	112.62	25.00	334.00	0.31	0.956	-0.210	0.030	-0.237

TABLE VIII. POLICY-WISE SUMMARY FOR FREEWAY: REWARD, LOW CONFIDENCE ACTION RATIO, AND TD ERROR

Policy	Reward				Low Conf 50%		Low Conf 10%	
	Mean	Min	Max	CV	Ratio (%)	TD Error	Ratio (%)	TD Error
softmax_0.1	21.37	19.00	26.00	0.07	0.000	0.000	0.000	0.000
softmax_0.5	21.36	19.00	26.00	0.07	0.003	-0.010	0.003	-0.010
softmax_1.0	22.00	18.00	27.00	0.08	0.568	-0.010	0.568	-0.010
softmax_2.0	20.30	15.00	25.00	0.07	8.492	-0.015	1.409	-0.019
softmax_10.0	3.82	0.00	9.00	0.42	0.017	-0.127	0.000	-0.043
deterministic	21.41	19.00	26.00	0.07	0.000	0.000	0.000	0.000
linear	18.92	13.00	24.00	0.10	1.823	-0.017	0.000	0.000
softmax_drop_1	22.06	17.00	27.00	0.08	0.573	-0.009	0.573	-0.009
softmax_drop_2	22.03	17.00	29.00	0.09	0.573	-0.011	0.573	-0.011
linear_drop_1	21.47	19.00	26.00	0.07	0.000	0.000	0.000	0.000
linear_drop_2	21.42	19.00	26.00	0.07	0.000	0.000	0.000	0.000

value function predicted. This observation is consistent with recent evidence that actor-critic value networks become miscalibrated on out-of-distribution state–action pairs, systematically over-estimating returns in precisely the regions we probe here [18]. A negative TD error ($\delta_t < 0$) means the experienced outcome, $R_{t+1}+\gamma V(S_{t+1})$, was significantly worse than what the value function $V(S_t)$ had predicted. The data provides unambiguous evidence for this. In Alien (Table IV), when the softmax_2.0 policy takes a low-confidence (50%) action, the resulting average TD error is -0.497. For the lowest-confidence (10%) actions, the error is -0.983. A similar pattern is observed in KFM (Table IX), where the softmax_2.0 policy yields a TD error of -6.111 and -10.671 on low-confidence actions of 50% and 10%, respectively. When the agent is forced to take an action it deems unlikely, the consequence is almost always worse than it expected.

3) Step 3: This is a clear symptom of value function overfitting. This chain of evidence points to a single, compelling conclusion: the agent's value function is overfit to the high-density regions of the policy distribution it experienced during training. During PPO's on-policy training, the critic (the value function) is trained to minimize TD error on trajectories generated by the current actor policy. This training data is naturally dominated by high-confidence actions. Consequently, the critic becomes an expert at predicting the value of these frequently observed, "good" state-action pairs. Conversely, it sees very few examples of low-confidence actions and their outcomes, leaving the value function undertrained and inaccurate for this part of the state-action space.

When we evaluate with a high-entropy policy, we are effectively sampling from low-density regions of the policy distribution. We force the agent to explore these infrequently

TABLE IX. POLICY-WISE SUMMARY FOR KFM: REWARD, LOW CONFIDENCE ACTION RATIO, AND TD ERROR

Policy	Reward				Low Conf 50%		Low Conf 10%	
	Mean	Min	Max	CV	Ratio (%)	TD Error	Ratio (%)	TD Error
softmax_0.1	31318.40	7000.00	60800.00	0.34	0.436	2.323	0.132	1.797
softmax_0.5	19289.60	8000.00	33000.00	0.23	2.507	-0.629	0.354	-2.660
softmax_1.0	13739.80	5800.00	24500.00	0.24	2.784	-7.744	0.537	-5.201
softmax_2.0	8113.20	2600.00	14900.00	0.31	6.274	-6.111	2.491	-10.671
softmax_10.0	966.20	0.00	2900.00	0.56	0.000	0.000	0.000	0.000
deterministic	22802.80	3400.00	51600.00	0.54	0.000	0.000	0.000	0.000
linear	2835.40	500.00	7100.00	0.41	11.982	-9.045	0.000	0.000
softmax_drop_1	13951.80	3800.00	22700.00	0.23	2.778	-7.110	0.546	-4.115
softmax_drop_2	14117.20	5200.00	26200.00	0.23	2.781	-7.109	0.543	-3.958
linear_drop_1	4168.20	700.00	12500.00	0.38	4.888	-8.184	0.770	-9.098
linear_drop_2	4468.60	900.00	18400.00	0.39	3.271	-7.312	0.627	-8.463

TABLE X. POLICY-WISE SUMMARY FOR PONG: REWARD, LOW CONFIDENCE ACTION RATIO, AND TD ERROR

Policy	Reward				Low Conf 50%		Low Conf 10%	
	Mean	Min	Max	CV	Ratio (%)	TD Error	Ratio (%)	TD Error
softmax_0.1	20.99	20.00	21.00	0.01	1.306	0.000	0.260	0.000
softmax_0.5	20.95	20.00	21.00	0.01	5.144	-0.000	0.609	-0.001
softmax_1.0	20.84	14.00	21.00	0.03	6.116	-0.001	0.489	-0.009
softmax_2.0	14.61	-19.00	21.00	0.38	6.004	-0.020	0.388	-0.143
softmax_10.0	-19.08	-21.00	-12.00	-0.08	0.920	-0.058	0.002	-0.071
deterministic	21.00	21.00	21.00	0.00	0.000	0.000	0.000	0.000
linear	-14.93	-21.00	-3.00	-0.22	4.203	-0.032	0.145	-0.041
softmax_drop_1	20.81	13.00	21.00	0.03	6.956	-0.001	3.134	-0.001
softmax_drop_2	20.82	14.00	21.00	0.03	9.339	-0.001	7.152	-0.001
linear_drop_1	-9.83	-20.00	15.00	-0.56	2.434	-0.040	0.084	-0.044
linear_drop_2	2.80	-19.00	21.00	2.90	1.639	-0.042	0.070	-0.059

TABLE XI. POLICY-WISE SUMMARY FOR QBERT: REWARD, LOW CONFIDENCE ACTION RATIO, AND TD ERROR

Policy	Reward				Low Conf 50%		Low Conf 10%	
	Mean	Min	Max	CV	Ratio (%)	TD Error	Ratio (%)	TD Error
softmax_0.1	3777.80	800.00	4175.00	0.24	0.800	2.531	0.171	-0.199
softmax_0.5	3439.85	575.00	4200.00	0.37	3.279	1.720	0.486	-0.235
softmax_1.0	2404.60	125.00	4450.00	0.68	4.653	-0.332	0.733	0.577
softmax_2.0	844.15	150.00	4200.00	0.81	7.433	-0.515	0.730	-0.257
softmax_10.0	312.35	0.00	1200.00	0.64	5.489	-2.509	0.000	0.000
deterministic	3894.00	800.00	4050.00	0.18	0.000	0.000	0.000	0.000
linear	371.55	25.00	1325.00	0.56	2.374	-0.304	0.083	-0.803
softmax_drop_1	2370.35	225.00	4600.00	0.69	4.625	-0.186	1.045	0.132
softmax_drop_2	2457.75	325.00	4200.00	0.67	4.711	-0.030	2.299	-0.477
linear_drop_1	481.15	0.00	4050.00	0.73	0.954	-0.911	0.020	-1.776
linear_drop_2	1201.20	125.00	4150.00	1.14	1.296	-1.239	0.028	0.054

sampled transitions, and the significant, negative TD errors are the direct result. The critic, having not been adequately trained on these inputs, provides overly optimistic value estimates and is then surprised by the poor outcomes. This is a classic symptom of statistical overfitting, where a model fails to generalize from its training distribution to a different test distribution. The success of deterministic evaluation is therefore not just a matter of exploitation; it is a strategy that implicitly hides this overfitting by constraining the agent to the well-modeled, high-density regions where its value function is accurate.

V. DISCUSSION

Our empirical analyses provide multifaceted insights into the consequences of employing stochastic evaluation policies for agents trained with PPO. The findings have significant practical implications for RL practitioners and open up new avenues for future research.

A. Synthesis of Findings: A Unified View

The narrative that emerges from our results is clear and consistent. The central finding of this study is that the value functions of PPO-trained agents are often overfitted to the high-confidence, on-policy actions experienced during training. This overfitting, compounded by the objective mismatch between a stochastic training goal and a deterministic evaluation goal, is revealed through the following mechanism:

1) Stochastic evaluation policies act as a diagnostic probe: By increasing policy entropy, we force the agent to sample low-confidence actions more frequently. This pushes the agent

TABLE XII. AVERAGE ENTROPY PER POLICY AND ENVIRONMENT

Policy	Alien	Boxing	Breakout	Enduro	Freeway	KFM	Pong	Qbert
softmax_0.1	0.216	0.057	0.183	0.195	0.000	0.206	0.198	0.112
softmax_0.5	0.998	0.289	0.482	0.820	0.000	1.058	0.803	0.544
softmax_1.0	1.577	0.575	0.727	1.234	0.036	1.529	1.132	0.871
softmax_2.0	2.144	1.138	0.987	1.592	0.335	1.958	1.396	1.193
softmax_10.0	2.820	2.761	1.329	2.137	1.042	2.585	1.756	1.692
deterministic	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
linear	2.677	2.699	1.015	1.905	0.514	2.367	1.416	1.501
softmax_drop_1	1.575	0.577	0.676	1.224	0.032	1.529	1.073	0.857
softmax_drop_2	1.572	0.575	0.495	1.209	0.000	1.527	0.974	0.796
linear_drop_1	2.593	2.575	0.565	1.806	0.000	2.163	1.202	1.279
linear_drop_2	2.517	2.475	0.000	1.704	0.000	2.112	0.895	0.888

to rely on its value function in low-density regions of the state-action space that were infrequently visited during training.

2) The probe reveals inaccurate and overly optimistic value estimates: The value function, having been poorly generalized in these regions, proves to be systematically over-optimistic. This inaccuracy is directly measured by the large, negative TD errors that result when these low-confidence actions are taken.

3) Inaccurate value estimates lead to performance degradation: This failure of the value function to accurately price the consequences of suboptimal actions is the direct cause of the observed performance degradation. The success of deterministic evaluation is therefore a strategy that succeeds primarily because it confines the agent to the well-modeled, high-density regions of its policy distribution where the value function is reliable. Conversely, the superior performance of low-to-moderate entropy policies in five of the eight environments (Alien, Breakout, Enduro, KFM, and Freeway) indicates that deterministic evaluation is not universally optimal. In these cases, the agent's performance is limited by factors other than value function overfitting. We hypothesize that modest stochasticity helps overcome challenges like perceptual aliasing (where different states appear identical) or helps the policy escape repetitive, suboptimal loops that a purely deterministic approach might get trapped in. This highlights a fundamental trade-off: while deterministic evaluation protects against the dangers of a poorly generalized value function, it may prevent the agent from achieving a higher score if the environment dynamics reward a degree of randomness.

B. Practical Implications for Reinforcement Learning Practitioners

Based on our findings, we propose the following guidelines for practitioners working with policy gradient algorithms:

1) Default to low-entropy evaluation (tune temperature): Deterministic (argmax) is a strong baseline, but not uniformly optimal across our Atari suite. Practitioners should treat evaluation entropy as a hyperparameter: begin with argmax or a low-temperature softmax, and select the setting that maximizes mean return and stability on held-out evaluation seeds.

2) Consider slightly stochastic evaluation in specific cases: In environments where there is a strong reason to suspect the presence of perceptual aliasing or where the policy might be trapped in a local optimum, evaluating with a small amount of stochasticity can be beneficial. As observed in Breakout, a low-temperature softmax policy (e.g., with temperature T=0.5) might allow the agent to escape repetitive patterns and achieve higher performance. In environments such as Breakout, we hypothesize that a purely deterministic policy may become trapped in a simple but less effective strategy. The evaluation temperature should be treated as a hyperparameter to be tuned based on the specific problem at hand.

3) Use TD error on low-confidence actions as a diagnostic tool: Our results highlight the diagnostic power of TD error. Monitoring the TD error on actions with low selection probability during the training process can serve as a powerful indicator of value function overfitting. A widening gap between the TD error on high-confidence versus low-confidence actions should be seen as a red flag, signaling a potential generalization problem.

C. Limitations and Future Work

While this study provides a comprehensive analysis, it has several limitations that point toward important directions for future research.

1) Algorithmic scope: Our investigation was confined to the PPO algorithm. It is an open question whether these findings generalize to other families of RL algorithms. Of particular interest are off-policy actor-critic methods, such as SAC, which explicitly optimize an entropy-regularized objective. Because an agent trained with SAC is optimized to act stochastically, it is plausible that such an agent would exhibit less performance degradation, or perhaps even an improvement, under stochastic evaluation. A comparative study is needed to resolve this.

2) Environment scope: The analysis was conducted on eight discrete-action Atari games. Future work should extend this investigation to other domains, such as continuous control tasks (e.g., MuJoCo benchmarks) and more complex, modern game environments, to assess the generality of our conclusions.

3) Exploring regularization: Our findings indicate that value function overfitting is a key limiting factor. This motivates future work to test empirically whether regularization techniques known to combat overfitting in supervised learning can enhance the value function's generalization in an RL context. Researchers and practitioners may apply methods such as weight decay, dropout, spectral normalization, or ensemble methods to the critic network during PPO training. An effective

regularization scheme should reduce the TD error on low-confidence actions and, as a result, diminish the performance gap between deterministic and stochastic evaluation.

VI. Conclusions

This paper has presented a rigorous empirical investigation into the common practice of training reinforcement learning agents with a stochastic policy but evaluating them deterministically. Our findings demonstrate that this practice is not merely a convention but is often a well-justified approach that mitigates a critical, underlying weakness in deep RL agents: the overfitting of the learned value function to its most frequently observed state-action pairs.

We have shown that higher evaluation-time entropy often degrades performance and stability—particularly in brittle environments—yet modest stochasticity can improve returns in several games (e.g., Alien, Breakout, Enduro, KFM, Freeway). Thus, evaluation entropy should be selected empirically, balancing exploitation against environment-specific benefits of limited randomness. The root cause of performance degradation, in cases where it occurs, is that stochastic policies force the agent to sample low-confidence actions, exposing it to low-density regions of the state-action space where its value function is demonstrably inaccurate and overly optimistic. This inaccuracy is a direct manifestation of overfitting to the on-policy training distribution, compounded by the inherent mismatch between the entropy-regularized stochastic objective used in training and the purely exploitative deterministic objective used in evaluation.

Ultimately, the choice of an evaluation policy transcends the simple dichotomy of exploration versus exploitation. It is a decision that interacts deeply with the generalization limits of the agent's learned models. Deterministic evaluation often succeeds because it acts as a safeguard, confining the agent to the high-density regions of its policy distribution where its internal value model is most reliable. However, our work also shows this is not a universal solution, as some environments reward modest stochasticity that can overcome other limitations of the learned policy. This work provides a clear empirical framework and a causal, TD-error-based mechanism for understanding this critical aspect of deep reinforcement learning applications, offering both practical guidance for practitioners and a diagnostic tool for future research aimed at building more robust and generalizable agents.

Acknowledgment

This research was supported by the 2024 KNUDP(Korea National University Development Project) funded by the Ministry of Education(MOE, Korea) and National Research Foundation of Korea(NRF)

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-Level Control through Deep Reinforcement Learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[2] A. Karalakou, D. Troullinos, G. Chalkiadakis, and M. Papageorgiou, "Deep Reinforcement Learning Reward Function Design for Autonomous Driving in Lane-Free Traffic," *Systems*, vol. 11, no. 3, 2023.

[3] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[4] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[5] R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, and M. Bellemare, "Deep reinforcement learning at the edge of the statistical precipice," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 29 304–29 320.

[6] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *International conference on machine learning (ICML)*, 2017, pp. 1352–1361.

[7] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning (ICML)*, 2018, pp. 1861–1870.

[8] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[9] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International Conference on Machine Learning (ICML)*, 2015, pp. 1889–1897.

[10] N. Milosevic, J. Müller, and N. Scherf, "Central Path Proximal Policy Optimization," in *The Exploration in AI Today Workshop at International conference on machine learning (ICML)*, 2025.

[11] C. B. Tan, E. Toledo, B. Ellis, J. N. Foerster, and F. Huszár, "Beyond the Boundaries of Proximal Policy Optimization," *arXiv preprint arXiv:2411.00666*, 2024.

[12] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International Conference on Machine Learning (ICML)*, 2016, pp. 1928–1937.

[13] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *International Conference on Machine Learning (ICML)*, 2014, pp. 387–395.

[14] M. Schilling, "Avoid Overfitting in Deep Reinforcement Learning: Increasing Robustness Through Decentralized Control," in *International Conference on Artificial Neural Networks (ICANN)*, 2021, pp. 638–649.

[15] E. Korkmaz, "A Survey Analyzing Generalization in Deep Reinforcement Learning," *arXiv preprint arXiv:2401.02349*, 2024.

[16] C. Zhang, O. Vinyals, R. Munos, and S. Bengio, "A Study on Overfitting in Deep Reinforcement Learning," *arXiv preprint arXiv:1804.06893*, 2018.

[17] T. Hongyao and B. Glen, "Improving Deep Reinforcement Learning by Reducing the Chain Effect of Value and Policy Churn," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[18] Q. Li, A. Kumar, I. Kostrikov, and S. Levine, "Efficient Deep Reinforcement Learning Requires Regulating Overfitting," in *International Conference on Learning Representations (ICLR)*, 2023.

[19] T. Zahavy, M. Haroush, N. Merlis, D. J. Mankowitz, and S. Mannor, "Learn what not to learn: Action elimination with deep reinforcement learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.

[20] K. Xue, X. Lin, Y. Shi, S. Kai, S. Xu, and C. Qian, "Escaping Local Optima in Global Placement," *arXiv preprint arXiv:2402.18311*, 2024.

[21] H.-R. Pan and B. Schölkopf, "Skill or Luck? Return Decomposition via Advantage Functions," in *International Conference on Learning Representations (ICLR)*, 2024.

Expert Systems in Tuberculosis Prevention Established in Certainty Factor

Inooc Rubio Paucar[1], Cesar Yactayo-Arias[2], Laberiano Andrade-Arenas[3]
Facultad de Ingeniería y Negocios, Universidad Privada Norbert Wiener, Lima, Perú[1]
Departamento de Estudios Generales, Universidad Continental, Lima, Perú[2]
Facultad de Ciencias e Ingeniería, Universidad de Ciencias y Humanidades, Lima, Perú[3]

***Abstract*—Tuberculosis remains a highly relevant public health concern, especially in contexts with limited access to medical services, highlighting the need for tools that support early diagnosis. In this study, a web-based expert system was developed to assist in tuberculosis detection, using Buchanan's methodology, which consists of five phases: identification, conceptualization, formalization, implementation, and validation. The system was designed with a knowledge rule-based approach and incorporated the Certainty Factor to quantify confidence in diagnostic conclusions. Validation was carried out through expert judgment using a 15-question survey. The results showed a high overall positive consensus, with question 13 standing out as it obtained the highest mean score (4.80) and the lowest dispersion (SD = 0.61), reflecting the most favorable perception and greatest agreement among the experts. Conversely, question 4 recorded the lowest mean score (4.00) and the highest dispersion (SD = 1.12), indicating aspects of the system that generated more divided opinions. Overall, these findings confirm that the system is effective, reliable, and usable, making it a relevant tool to support clinical decision-making in resource-limited settings. As an additional contribution, the integration of complementary technologies is suggested, such as (ML) algorithms, radiological image analysis, and mobile applications for symptom tracking, in order to optimize early detection and strengthen clinical care for tuberculosis.**

***Keywords*—Buchanan's methodology; certainty factor; expert system; public health; tuberculosis; web application**

I. Introduction

Tuberculosis (TB) is an infectious disease caused by Mycobacterium tuberculosis, which primarily affects the respiratory system, although it can also compromise other vital organs. Since ancient times, TB has represented a significant threat to global public health and, despite advances in research, diagnosis, and therapies, it continues to be one of the major challenges in the healthcare field. According to the World Health Organization (WHO), millions of new cases and hundreds of thousands of deaths are reported each year, keeping TB among the deadliest infectious diseases worldwide, especially in low-resource regions [1] [2]. Likewise, it is necessary to implement preventive measures that allow anticipating the onset of symptoms and, in this way, facilitate the administration of timely and appropriate treatment for each person.

The control of tuberculosis faces various structural and clinical obstacles. One of the most critical is the difficulty in achieving early and accurate diagnosis, since initial symptoms such as persistent cough, mild fever, or weight loss are often confused with other common respiratory diseases. This delay in detection not only prevents the timely administration of treatment but also facilitates community transmission of the disease. In addition, the lack of knowledge among the population about the modes of transmission, prevention, and treatment of TB intensifies its spread, particularly in environments where overcrowding, poverty, and limited healthcare coverage persist [3]. Added to this is the emergence of drug-resistant strains (MDR-TB and XDR-TB), which hinder therapeutic effectiveness, increase healthcare costs, and prolong treatment regimens. Finally, insufficient medical infrastructure and a lack of specialized personnel in many underdeveloped regions limit diagnostic, monitoring, and control processes, weakening eradication programs established by international organizations [4]. Altogether, these limitations make it evident that traditional approaches are insufficient to address the magnitude and complexity of the problem.

In light of this reality, there is a need to incorporate innovative technological tools to support healthcare systems in the identification, control, and treatment of tuberculosis. In this sense, expert systems and (AI)represent an alternative with great potential, as they allow clinical information to be processed under conditions of uncertainty, reduce the margin of error in medical decision-making, and generate faster and more reliable diagnoses [5]. Moreover, these tools have the capacity to optimize available resources, as they can be implemented in low-infrastructure environments and simultaneously serve a large number of users. Their usefulness lies not only in improving early diagnosis but also in supporting less-experienced physicians, standardizing clinical criteria, and providing support in high-demand healthcare contexts. The application of these technologies can significantly contribute to strengthening public health policies, improving the operational efficiency of health services, and ultimately reducing morbidity and mortality rates associated with tuberculosis [6] [7]. Thus, the development and implementation of expert systems is justified as a key strategy not only for individual clinical care but also for the global epidemiological control of this disease.

Therefore, this research focuses on developing an expert system (ES) for the early identification of tuberculosis, applying Buchanan's methodology for the construction of expert systems. The objective of the study is to diagnose tuberculosis based on symptoms reported by the patient and to provide a level of certainty through the use of the Certainty Factor (CF), with the purpose of supporting medical personnel in clinical decision-making and contributing to the prevention and control of tuberculosis.

II. Literature Review

This section presents the structure of the literature review, which is organized around two main axes. First, related works are addressed, where previous research by various authors who have applied expert systems in the medical field is analyzed. The purpose of this section is to contextualize the object of study, identifying advances and limitations in previous research in order to provide a broader and more well-founded perspective on the topic. Second, the theoretical foundations are developed, presenting the conceptual frameworks and theories related to the central variables of this research: Expert Systems and Tuberculosis.

A. Related Work

Various studies have contributed to advances in the detection of tuberculosis. In particular, the authors [8] [9] [10] highlight the importance of web-based expert systems to facilitate the diagnosis of respiratory diseases. Their proposal focuses on six pathologies: tuberculosis, bronchitis, pneumonia, pleurisy, tonsillitis, and asthma. To achieve more accurate results, the system uses 40 symptomatic indicators as a source of information. The reported results show that the application reaches a considerable level of effectiveness, offering reliable diagnoses for these diseases. In order to enrich previous studies, an ES was designed exclusively focused on tuberculosis diagnosis, especially aimed at people who do not have timely access to medical checkups. For this purpose, the Case-Based Reasoning (CBR) method was applied, which allows comparing the symptoms of a new patient with records of previous cases and thus estimating the most probable diagnosis. This approach, supported by expert knowledge, is presented as a practical and reliable alternative within the healthcare field. The results showed that pulmonary tuberculosis reached 85% similarity, while the extrapulmonary form obtained 62%, which allowed concluding that the analyzed case corresponded to pulmonary tuberculosis. In hospital settings, limitations persist in guaranteeing quality and truly personalized care for each patient. In response to this challenge, the developed ES is presented as a clinical support tool that follows a specialized assistance pattern in cases of pulmonary tuberculosis, even before confirmation through laboratory tests. Its architecture is based on a rule base and a patient knowledge base, which interact with an inference engine responsible for processing the data entered through the user interface. The usability evaluation yielded an average score of 5.6 out of 7, reflecting a performance above the expected standard. In this sense, the Tuberculosis Diagnosis ES, developed with Exsys Corvid, constitutes a robust and effective technological solution, rule-based, aimed at optimizing the diagnostic process and improving the quality of hospital services [11] [12]. Thus, a development based on expert systems and information technology aims to optimize TB diagnosis, overcoming existing limitations in timely access to healthcare services. For this purpose, an application supported by expert system architecture was implemented, applying the Forward Chaining method, which enables the processing of clinical data and symptoms to generate diagnostic inferences. The result of this work was an application capable of identifying TB accurately and providing users with structured information on symptoms, as well as recommendations and alternative solutions to the disease, consolidating itself as an efficient technological tool to support the diagnostic process and medical guidance [13] [14].

Tuberculosis is an infectious disease mainly of bacterial origin that critically affects the respiratory system and, in some cases, other organs, making it a highly relevant public health problem. In this context, various studies have proposed technological solutions aimed at improving its diagnosis and surveillance. For example, one study aimed to design an ES that replicated the clinical reasoning of a health professional to support the diagnosis of pulmonary tuberculosis based on the identification of symptoms in patients, considering economic limitations in access to healthcare services. This system, based on Bayes' Theorem formulated by Thomas Bayes and refined by Laplace, achieved a positive diagnostic rate of 87%, demonstrating its effectiveness in the early detection of the disease and its applicability in resource-limited settings [15] [16]. Complementarily, a web-based ES was developed that used the Dempster-Shafer inference method, integrating a PHP backend with MySQL and a frontend built with HTML, CSS, and JQuery. This project followed a structured methodology that included preliminary studies, literature review, symptom collection, knowledge rule design, modeling, implementation, and evaluation, achieving the generation of reliable probabilistic diagnoses for both healthcare professionals and general users. In a different line, another work proposed an online ES oriented to the epidemiological surveillance of tuberculosis, developed with web platforms and the Swi-Prolog programming language, following Buchanan's methodology composed of four phases. Validation, through a 15-item questionnaire distributed across three dimensions, reported an 85% acceptance rate among users, supporting its relevance and effectiveness as a support tool for tuberculosis prevention and awareness [17] [18]. In another instance, [19] [20] explored the application of ML techniques in the diagnosis of tuberculosis, evaluating the CAD4TB computer-assisted detection system, originally designed for adults, on pediatric radiographs of children under 13 years of age. The model, trained with pediatric radiological images and validated by specialists through a prospective observational methodology, showed a significant improvement in its performance after fine-tuning, increasing from an AUC of 0.58 to 0.72 ($p = 0.0016$), evidencing the potential of this tool to strengthen diagnosis in pediatric populations and highlighting the importance of continuing validation in larger and more diverse samples.

Likewise, the authors [21] [22] developed an expert API for the early detection of tuberculosis, implementing inference algorithms such as Forward Chaining and Certainty Factor in order to support preliminary diagnosis of the disease. The proposal materialized through an API-based architecture that processes symptoms and generates diagnostic inferences, while its validation was carried out through functionality tests and in-depth interviews with medical specialists. Satisfactory results were obtained, evidencing that the system is capable of providing reliable support in the early identification of tuberculosis, thus constituting a viable technological tool to strengthen prevention and diagnostic tasks in healthcare. In another study, a Bayesian inference-based ES was developed to facilitate and speed up tuberculosis diagnosis, reducing the time required by conventional procedures that include clinical interviews and sputum laboratory analysis. Microsoft

Visual Studio 2010 was used as the programming platform and Microsoft Access 2010 as the data management system for its construction. The methodology implemented included the creation of a medical knowledge base, the integration of a Bayes-based inference engine, and the validation of the system through tests with real clinical cases. The results showed that the tool is capable of analyzing patient symptoms and clinical data to calculate the probability of tuberculosis, constituting a reliable resource that optimizes diagnosis and supports medical decision-making, offering advantages over traditional methods in both efficiency and accuracy [23] [24] [25]. Thus, the development of a hybrid Clinical Decision Support System (CDSS) aimed at recommending individualized treatments for patients with drug-resistant tuberculosis facilitated the application of personalized therapies by less experienced physicians. A hybrid approach combining an expert knowledge base with ML was used to identify complex patterns in clinical data, evaluating its performance with metrics such as precision at 1, mean reciprocal rank, and mean average precision. Indeed, the methodology included the construction of the knowledge base, the iterative generation of training datasets, analysis through ML algorithms, and validation with independent data to detect overfitting. Finally, the results showed an accuracy of 95% in recommending the first treatment, although independent validation reduced accuracy to 78%, evidencing that this hybrid approach constitutes a promising step toward the automation of personalized treatment and can be applied to other pathologies [26] [27].

According to study [28], computers constitute a widely used tool in the medical field to support the diagnosis of respiratory diseases, including tuberculosis. In this context, the research focused on the development of an ES aimed at the early diagnosis of this pathology. This system estimates the probability level of a patient having the disease, expressed through a certainty value. This value is determined using the CF, a methodology that allows quantifying the degree of confidence in environments characterized by high uncertainty and that is also capable of representing absolute convictions within the diagnostic inference process. Taking the context into account, this research is complemented by the evaluation of an ES applied to ML techniques for the classification of TB bacteria through images. Additionally, algorithms such as Decision Trees, Random Forest (RF), and Adaptive Boosting (AdaBoost) were trained using a set of 126 images classified into two classes, as part of the methodology. As a result, it was observed that the Random Forest model was the most outstanding, reaching an accuracy of 85%, making it the most effective model for tuberculosis diagnosis through the ES [29] [30]. On the other hand, [31] [32] carried out a study focused on tuberculosis prevention through the implementation of a Bayes-based ES, with the aim of providing a more accurate diagnosis. The methodology included the identification of medical problems related to the treatment of the disease. It was observed that, in one patient, a high probability of tuberculosis could be determined, with a confidence value of 64%, evidencing that this approach enables medical personnel to diagnose and recognize symptoms quickly and efficiently. According to the authors [33] [34], the rapid spread of tuberculosis is favored by the lack of public knowledge about the disease. In this context, the implementation of an ES allows the application of early medical diagnostic techniques, with the aim of facilitating the prompt administration of treatment. Regarding the results obtained through system testing, together with validation using the black-box technique, it was demonstrated that each functionality operates correctly and that the information provided is reliable. Additionally, the system showed efficient performance in managing a high number of concurrent users, making it suitable for clinical environments with high demand for consultations.

In this sense, one of the deficiencies identified in previous studies is the lack of integration of web-based expert systems that combine rule-based inference, the certainty factor, and validation through expert judgment, which limits the applicability of findings in resource-limited contexts. Moreover, most studies focus on isolated methods, such as Bayesian inference or ML models, without combining multiple approaches, which reduces the generalization and robustness of the diagnostic support provided.

B. Theoretical Bases

1) Tuberculosis: TB is a chronic infectious disease primarily caused by Mycobacterium tuberculosis, an acid–alcohol-resistant bacillus with a high capacity for airborne transmission. Its most common form affects the respiratory system, although in extrapulmonary presentations it can involve various organs and tissues. Likewise, infection occurs after the inhalation of contaminated droplets that reach the pulmonary alveoli, where the microorganism establishes a complex interaction with the host's immune response. From a clinical and epidemiological perspective, TB represents a major global public health challenge due to its high impact in terms of prevalence and mortality, in addition to its close association with social determinants such as poverty, malnutrition, and overcrowding. Despite the availability of effective therapies, poor treatment adherence promotes the emergence of multidrug-resistant strains, which significantly complicates its control and eradication [35]. In summary, this disease causes severe pulmonary and respiratory complications that, over time, can significantly compromise both quality of life and life expectancy.

a) Pulmonary tuberculosis: It is the most common form of tuberculosis, characterized by infection of the lungs with Mycobacterium tuberculosis. It is transmitted primarily through respiratory droplets expelled when coughing, sneezing, or talking, making it the main source of infection in the community [36]. Clinical manifestations include chronic cough, hemoptysis, chest pain, fever, night sweats, weight loss, and fatigue.

b) Extrapulmonary tuberculosis: It refers to all forms of tuberculosis that affect organs and systems outside the respiratory system. It accounts for approximately 20% of cases and can involve lymph nodes, pleura, bones, meninges, kidneys, digestive system, skin, among others. Its symptoms depend on the affected site, and although it is not usually transmissible, it poses a clinical and diagnostic challenge due to the variety of presentations and associated complications [37].

Fig. 1 shows a horizontal bar graph representing the percentage of TB cases by age group during the years 2019, 2021, and 2022, highlighting that the 30 to 49 age groups

concentrate the highest proportion of cases, while the age extremes have a lower incidence. On the other hand, Fig. 2 presents a map of Peru with the TB incidence rate per 100,000 inhabitants in 2022, where the red regions show the highest risk (≥ 76), the yellow ones intermediate risk (50 to 74), and the green ones lower risk (< 50). This figure is complemented by a table detailing the cases by department and year, showing that departments such as Callao, Lima, and Ucayali have had the highest rates in recent years. Together, these images allow us to visualize both the age distribution and the geographic distribution of TB in the country over time.

2) Expert systems: An ES is a type of computational system within the field of AI designed to emulate the reasoning process and decision-making ability of a human specialist in a specific domain. Its operation is based on a knowledge base, which contains facts and rules about the application area, and an inference engine, responsible for applying these logical rules to solve problems, provide diagnoses, or recommend solutions. On the other hand, these systems do not seek to replace the human expert but rather to assist in solving complex problems, especially in contexts where specialized expertise is limited or unavailable. Moreover, their main value lies in offering consistency in reasoning, speed in generating solutions, and the possibility of transferring expert knowledge to technological platforms [39] [40] [41]. Ultimately, expert systems aim to emulate the knowledge of a human specialist in a given field in order to solve complex problems, provide recommendations, and support decision-making in an efficient and reliable manner, even in the absence of a human expert. Consequently, Fig. 3 shows the architecture of Buchanan's methodology for an expert system, illustrating the flow of knowledge from a human expert to an end user.

III. Methodology

A. Definition of the Buchanan Methodology

Buchanan's methodology constitutes a systematic approach applied to the analysis and design of systems, aimed at solving complex problems through an orderly, structured, and progressive process. Its foundation lies in breaking down the problem into sequential phases that enable progress from an abstract understanding of the need to the achievement of an implemented and operational solution. These phases include: identification, where the problem and its boundaries are recognized; conceptualization, in which objectives are defined and a preliminary conceptual model is established; formalization, which specifies the technical and functional requirements of the solution; implementation, where the design is materialized into a functional system; and execution or operation, the stage in which the solution is integrated into the real environment with monitoring and feedback mechanisms. This framework is illustrated in Fig. 4, which presents the phases of Buchanan's methodology, visually showing each stage of the process [42]. Consequently, this methodology reduces ambiguity, optimizes decision-making, and ensures consistency across stages, making it a robust tool to guarantee the quality, relevance, and effectiveness of developed systems.

1) Identification: In accordance with the established framework, this stage marks the beginning of Buchanan's methodological process, starting with the identification phase. Table I presents the Identification phase corresponding to Buchanan's methodology, in which the main issue is recognized: the difficulty of diagnosing tuberculosis early and accurately due to the similarity of its symptoms with other respiratory diseases. Based on this, the need is established to develop an expert system that supports medical personnel in clinical decision-making. Likewise, the objective of this phase is defined as the clear delimitation of the problem and the initial scope of the system, which are oriented toward the collection and analysis of symptoms through the CF model to provide a level of confidence in the diagnosis. Consequently, the expected benefit is highlighted: to contribute to the prevention and control of tuberculosis through a reliable computational tool that supports clinical diagnosis. In this sense, Table II represents the main sources of knowledge integrated into the expert system. These include specialists such as pulmonologists and infectious disease experts, who contribute to the identification of symptoms and risk factors, as well as official WHO guidelines that serve as references in the formulation of rules [43]. It is important to emphasize that the beginning of this phase constitutes the starting point for the development of subsequent processes.

2) Conceptualization: In this phase, the functional objectives and the conceptual model of the expert system are defined. The purpose is to establish an overall vision of how the proposed solution will operate, identifying its main components and the relationships among them, without yet delving into the technical details of implementation. Consequently, Fig. 5 describes the system objectives, outlining what it does allowing the user to enter symptoms, calculate a certainty factor, and generate recommendations—and what it does not do such as replacing a physician or laboratory tests. In addition, the diagram illustrates the conceptual model, which includes the key symptoms, how knowledge is represented (through rules and a certainty factor), and the definition of conceptual rules with examples [44]. Finally, it shows the expected outcome: an expert system report that, based on the certainty level, identifies the type of tuberculosis (active, extrapulmonary, or latent) and suggests actions such as visiting a healthcare center for confirmatory tests.

Knowledge Representation

The expert system's knowledge is represented by production rules. The structure of these rules is shown below:

$$\textbf{IF } (\textit{Condition}) \implies \textbf{THEN } (\textit{Conclusion}) \ [\textit{Certainty Factor}]$$

This formalism allows the system to combine multiple pieces of evidence and generate diagnostic conclusions with associated degrees of certainty. The CF ranges from 0 (no confidence) to 1 (absolute confidence), providing a way to handle uncertainty in medical decision-making.

3) Formalization: The formalization stage consists of transforming the knowledge obtained from experts and structured documentary sources into a format that can be processed by the system. Accordingly, Table IIIc presents the formalization of the expert system rules for tuberculosis diagnosis, organized into three subtables. Subtable III presents the rules based on clinical symptoms, which consider

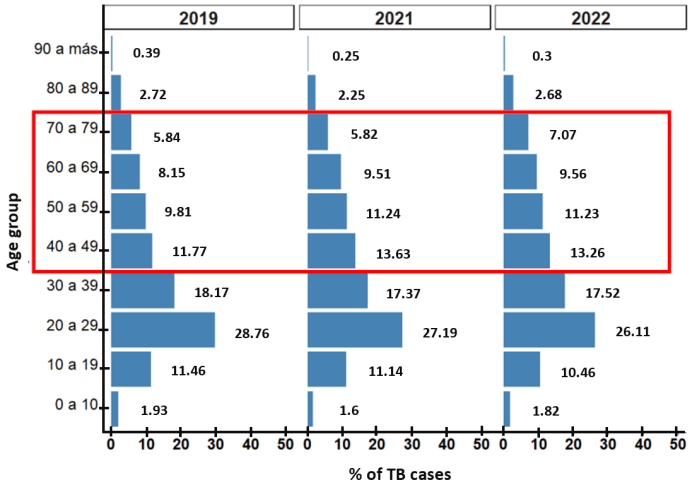
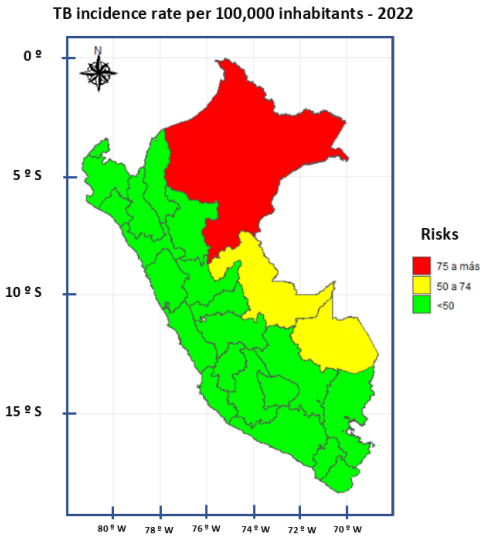


Fig. 1. Statistical graph of TB cases by age [38].



Departament	2019	2020	2021	2022
Peru	105.14	78.06	82.25	34.32
AMAZONAS	18.97	15.93	21.36	10.94
ANCASH	66.31	49.21	47.82	19.59
APURIMAC	20.35	12.77	16.38	9.77
AREQUIPA	60.71	44.81	46.00	17.24
AYACUCHO	31.87	23.50	36.23	10.72
CAJAMARCA	15.81	11.01	13.87	6.46
CALLAO	180.46	124.71	127.39	51.12
CUSCO	35.50	28.00	27.18	9.47
HUANCAVELICA	19.64	13.96	14.27	5.41
HUANUCO	60.96	48.80	54.83	27.80
ICA	161.25	113.72	112.13	44.90
JUNIN	59.50	42.67	45.93	19.86
LA LIBERTAD	85.36	59.75	64.93	29.99
LAMBAYEQUE	89.26	47.15	55.18	22.11
LIMA	131.09	109.57	115.28	53.86
LIMA CENTRO	264.66	173.99	181.92	76.62
LIMA ESTE	148.57	114.27	108.47	41.05
LIMA NORTE	125.11	94.53	97.39	41.81
LIMA SUR	104.96	81.66	81.62	29.47
LORETO	140.10	121.45	144.18	75.22
MADRE DE DIOS	191.10	129.45	152.93	52.29
MOQUEGUA	83.00	76.79	68.09	25.05
PASCO	12.15	11.40	9.89	4.83
PIURA	32.70	21.73	27.71	11.41
PUNO	28.88	23.83	31.93	9.78
SAN MARTIN	41.48	31.57	33.78	16.00
TACNA	131.17	111.33	99.05	42.67
TUMBES	54.65	55.66	37.23	23.50
UCAYALI	209.77	144.12	164.35	74.32

Fig. 2. TB cases by province per year [38].

combinations of manifestations such as persistent cough, fever, night sweats, or weight loss, each associated with a specific CF. Subtable IIIb includes the rules related to risk factors, such as recent exposure to the disease, HIV coinfection, or residence in endemic areas, which increase the diagnostic probability. In this regard, subtable IIIc contains the rules derived from diagnostic tests, such as chest X-rays, sputum smear microscopy, molecular tests, or cultures, which present the highest certainty values, highlighting their relevance in confirming the disease. On the other hand, Fig. 6 represents the database model, which is responsible for the flow and integrity of the data within the expert system [45]. In this sense, the TBExpertSystem database was implemented to comprehensively manage the information that feeds the expert system for tuberculosis diagnosis, in accordance with the business rules defined in the previous stages.

IV. Mathematical Formulas in Expert Systems

The following formulas are frequently used in expert systems for formalizing rules, handling uncertainty, and applying fuzzy logic. Therefore, mathematical expressions have been compiled that demonstrate the logic underlying the operation of these systems [46] [47].

1) Certainty Factor (CF): The CF is a mechanism that allows us to quantify the degree of confidence in a hypothesis when considering new evidence. This process adjusts the prior certainty of H based on the strength of the evidence E [see Eq. (1)].

$$CF_{H,E} = CF_H + CF_E \cdot (1 - CF_H) \tag{1}$$

Used to combine the certainty level of a hypothesis H with the evidence E.

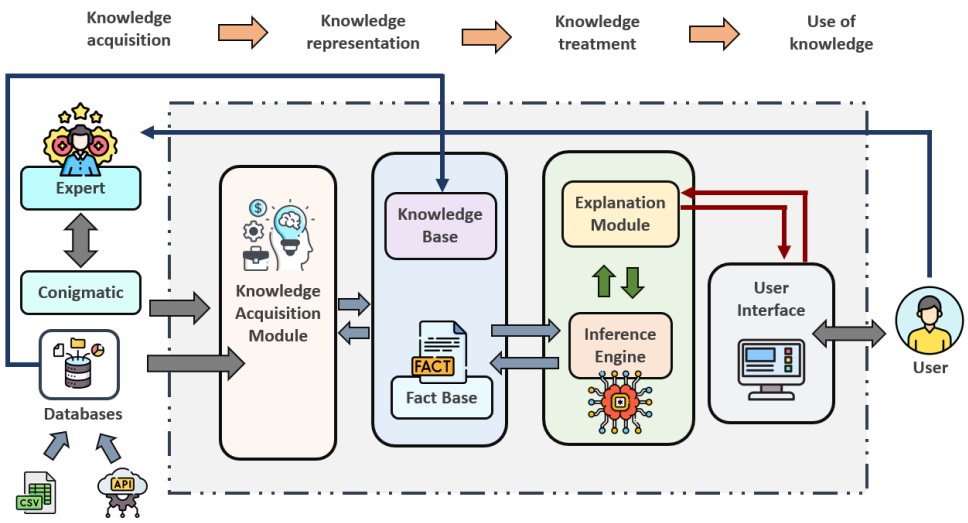


Fig. 3. Buchanan methodology architecture.

TABLE I. IDENTIFICATION PHASE: BUCHANAN METHODOLOGY

Element	Description
Identified problem	Difficulty in the early and accurate detection of tuberculosis because its symptoms can be confused with other respiratory diseases.
Need	Develop an expert system that supports medical personnel in clinical diagnosis, reducing ambiguity and improving decision-making.
Phase objective	Recognize and define the problem to be solved, establishing the initial scope of the system to be implemented.
Initial Scope	The system will be designed to collect patient-reported symptoms, process them using the CF model, and provide a level of certainty regarding the possible presence of tuberculosis.
Expected Benefit	Contribute to the prevention and control of tuberculosis by providing reliable computational support in the diagnostic process.

TABLE II. KNOWLEDGE SOURCES FOR THE EXPERT SYSTEM

Source	Description
Expert interviews	Consultations with pulmonologists and infectious disease specialists to determine the most relevant TB symptoms and risk factors.
WHO guidelines	Use of the Global Tuberculosis Report and official protocols as reference for rule formulation.
Clinical literature	Medical textbooks and peer-reviewed journals on internal medicine and infectious diseases.
Patient case data	Historical clinical cases used to validate CF and refine the knowledge base.

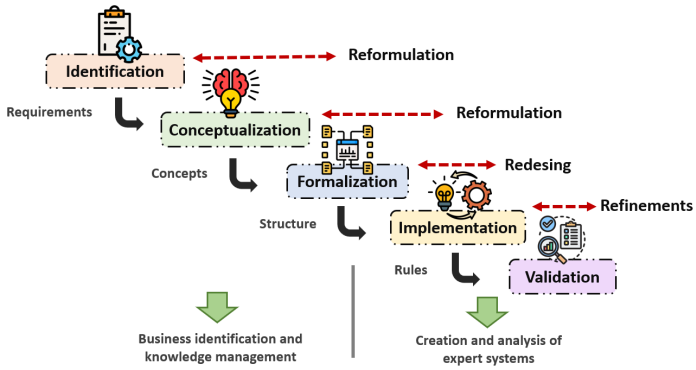


Fig. 4. Buchanan methodology.

2) Bayesian probability: Bayesian probability updates the degree of belief in a hypothesis H when new evidence E is observed. This process combines the prior probability with the likelihood of the evidence to obtain a more accurate estimation [see Eq. (2)].

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)} \tag{2}$$

Fundamental formula for probabilistic expert systems. It calculates the probability of hypothesis H given that evidence E occurred.

3) Bayesian probability with multiple evidences: This extended Bayesian approach calculates the probability of a hypothesis H when multiple evidences are considered simultaneously. It combines prior knowledge with the likelihood of each evidence to refine the overall estimation [see Eq. (3)].

$$P(H \mid E_1, E_2, \ldots, E_n) = \frac{P(H) \cdot \prod_{i=1}^{n} P(E_i \mid H)}{P(E_1, E_2, \ldots, E_n)} \tag{3}$$

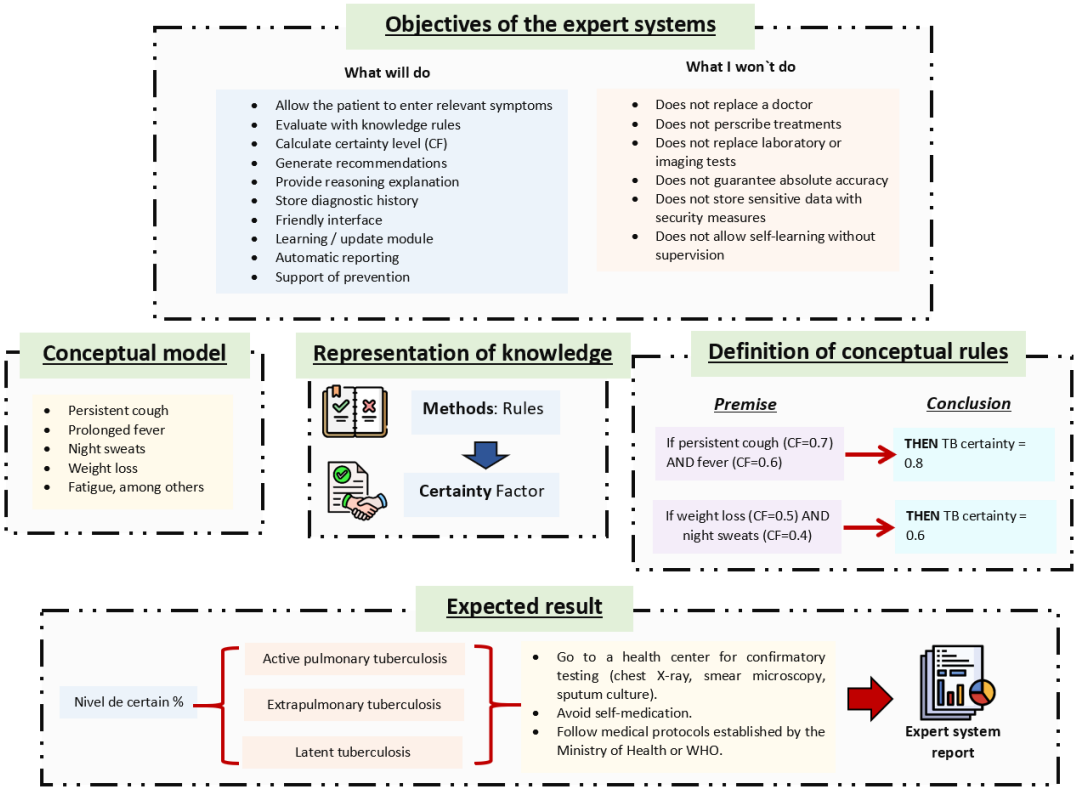


Fig. 5. Expert system process functionality.

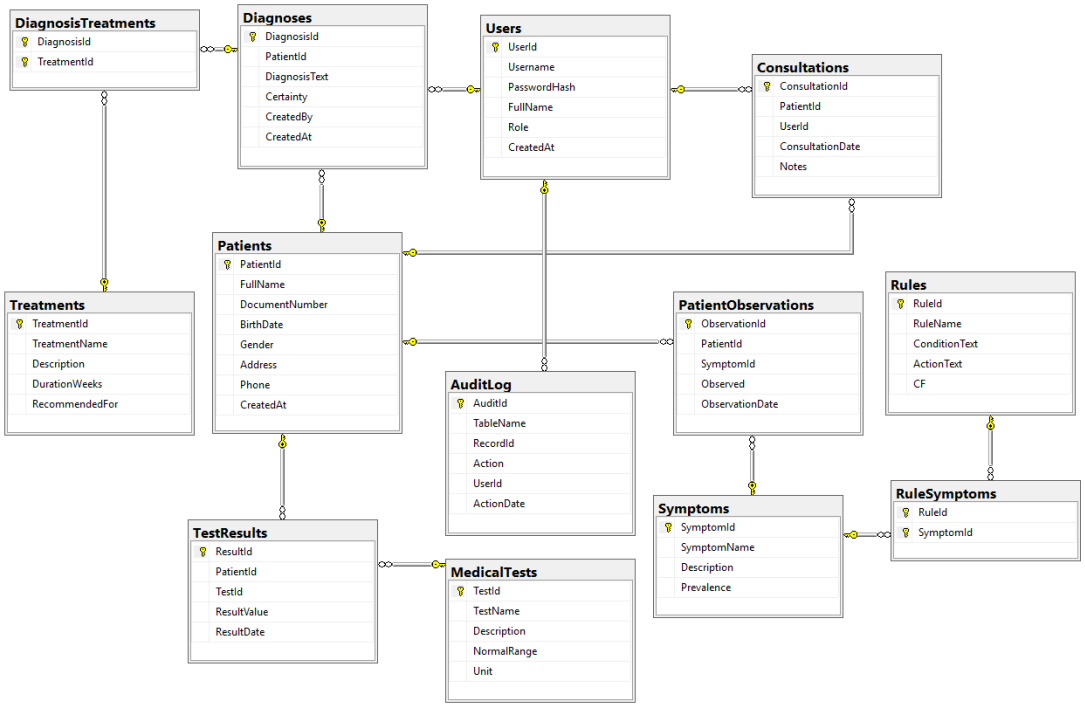


Fig. 6. Data flow for the expert system.

TABLE III. FORMALIZATION OF RULES FOR TUBERCULOSIS DIAGNOSIS

(A) RULES BASED ON CLINICAL SYMPTOMS

Rule	Condition	Certainty (CF)
R1	Persistent cough $\wedge$ Fever $\rightarrow$ TB	0.4
R2	Night sweats $\wedge$ Weight loss $\rightarrow$ TB	0.3
R3	Persistent cough $\wedge$ Hemoptysis $\rightarrow$ TB	0.5
R4	Chest pain $\wedge$ Breathing difficulty $\rightarrow$ TB	0.25
R5	Fatigue $\wedge$ Weakness $\rightarrow$ TB	0.2
R6	Fever $\wedge$ Night sweats $\rightarrow$ TB	0.35
R7	Persistent cough $\wedge$ Loss of appetite $\rightarrow$ TB	0.3
R8	Persistent cough $\wedge$ Weight loss $\rightarrow$ TB	0.45
R9	Chronic cough lasting $>$ 3 weeks $\rightarrow$ TB	0.55

(B) RULES BASED ON RISK FACTORS

Rule	Condition	Certainty (CF)
R10	Recent TB exposure $\wedge$ Persistent cough $\rightarrow$ TB	0.6
R11	HIV-positive patient $\wedge$ Fever $\wedge$ Night sweats $\rightarrow$ TB	0.65
R12	Living in high-risk area $\wedge$ Chronic cough $\rightarrow$ TB	0.5
R13	History of untreated TB contact $\rightarrow$ TB	0.55

(C) RULES BASED ON DIAGNOSTIC TESTS

Rule	Condition	Certainty (CF)
R14	Positive Mantoux test $\rightarrow$ TB	0.7
R15	Abnormal chest X-ray $\rightarrow$ TB	0.8
R16	Positive sputum smear microscopy $\rightarrow$ TB	0.85
R17	Positive GeneXpert test $\rightarrow$ TB	0.9
R18	Positive sputum culture test $\rightarrow$ TB	0.95

Extended version for multiple simultaneous pieces of evidence. It evaluates the probability of hypothesis H given evidences $E_1, E_2, \ldots, E_n$.

4) Fuzzy membership function (triangular): The triangular membership function represents how a value x belongs to a fuzzy set A with a gradual transition. It is widely used in fuzzy logic due to its simplicity and ability to approximate uncertainty smoothly [see Eq. (4)].

$$\mu_A(x) = \begin{cases} 0 & x \leq a, \\ \dfrac{x-a}{b-a} & a < x \leq b, \\ \dfrac{c-x}{c-b} & b < x < c, \\ 0 & x \geq c \end{cases} \tag{4}$$

Defines the membership of a value x to a fuzzy set A. The shape is triangular, which is commonly used in fuzzy logic.

5) Fuzzy inference (Mamdani rule): The Mamdani inference rule models the reasoning process in fuzzy systems by combining fuzzy sets through min–max operations. It determines the degree to which an output z belongs to set C given the inputs x and y [see Eq. (5)].

$$\mu_C(z) = \sup_{x,y} \left(\min\left(\mu_A(x), \mu_B(y)\right)\right) \tag{5}$$

Represents a fuzzy inference for a rule. If $x \in A$ and $y \in B$, then $z \in C$.

6) Defuzzification by centroid: The centroid defuzzification method transforms a fuzzy output set into a crisp numerical value. It is the most common technique in fuzzy control systems, as it provides a balanced representation of all possible outcomes [see Eq. (6)].

$$z^* = \frac{\int_Z z \cdot \mu_C(z)\, dz}{\int_Z \mu_C(z)\, dz} \tag{6}$$

Converts a fuzzy value into a numerical output. This method is widely used in fuzzy control systems.

7) Information entropy: Information entropy quantifies the level of uncertainty or randomness in a variable X. It measures how much information is gained or how unpredictable the outcomes are within a system [see Eq. (7)].

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{7}$$

Measures the uncertainty associated with a random variable X. Entropy quantifies the amount of information or unpredictability.

a) Implementation: Finally, the implementation phase corresponds to the translation of the previous stages into a web-based expert system, in which the knowledge base, inference engine, and user interface are integrated into a functional platform. In this stage, the defined rules are materialized, the reasoning mechanisms are programmed, and the functionalities that enable interaction with the end user

are designed [48]. Thus, Fig. 7 presents the user interface corresponding to the login module, which allows users to enter their credentials (username, password, and assigned role). In addition, it incorporates a search module that verifies whether the user belongs to the healthcare system, ensuring real-time data validation. Finally, a registration form is provided to enable the creation of new accounts, ensuring proper access management and traceability of user profiles within the web expert system.

In another instance, Fig. 8 presents three user interface panels of an information or expert system on tuberculosis (TB). The panel in Fig. 8a, labeled "Administrative panel," is designed for user management, displaying a list of records, search options, and buttons to register, update, or delete user codes. The panel in Fig. 8b is a "Rules Record," where detection rules for different types of TB are defined and managed, including details such as rule name, condition, action, and centrality. Finally, the panel in Fig. 8c is the "User panel," which shows the medical history of a patient (Juan Alvarez Perez) and a "Quick Symptom Form" that allows the user or physician to register symptoms and predict the type of TB, highlighting the latest medical result and the status of the disease.

V. Results

This section details the findings of the statistical analysis of the user perception scores assigned to the 15 interface prototypes, a fundamental process for validating the design and functionality of the web-based expert system by healthcare professionals. The evaluation was based on a 15-question questionnaire structured around five key dimensions (Usability, Technical Functioning, Knowledge Base, Inference Engine, and Expected Impact), as detailed in Table IV. Fig. 9 displays the distribution of responses (on a scale of 1 to 5) and showed a high overall positive consensus, with the majority of scores concentrated at levels 4 and 5. Additionally, the descriptive metrics presented in Table V confirm the consistency of the sample (N=20 per prototype). Prototype 13 (P-13) stood out with the highest mean (4.80) and the lowest dispersion (SD = 0.61), indicating the most favorable perception and the greatest agreement among experts. In contrast, Prototype 4 (P-4) recorded the lowest mean (4.00) and the greatest dispersion (SD = 1.12), indicating the area with the most divided opinions.

VI. Discussion

This section discusses the findings obtained from the evaluation of the web expert system prototypes for tuberculosis diagnosis, relating them to previous studies and evidence reported in the literature. It analyzes healthcare professionals' perceptions regarding the system's usability, technical performance, knowledge base, and inference engine, as well as its relevance as a clinical support tool. Furthermore, it examines how inference methods and knowledge structuring contribute to optimizing the early detection of tuberculosis. This discussion contextualizes the results, highlights the original contributions of the study, and establishes its relevance in the field of public health. The findings from the analysis of the 15 web expert system prototypes for tuberculosis diagnosis generally show a high level of positive consensus among healthcare professionals, with Prototype 13 standing out (mean = 4.80; SD = 0.61), indicating both favorable perceptions and a high level of agreement among experts. These results are consistent with previous studies that have demonstrated the effectiveness of rule-based expert systems and specialist knowledge in supporting clinical decision-making in respiratory diseases [11] [9] [10]. Likewise, the evaluation through a structured questionnaire covering five dimensions (Usability, Technical Performance, Knowledge Base, Inference Engine, and Expected Impact) made it possible to systematically identify strengths and areas for improvement, following strategies similar to those reported in the validation of web expert systems in other clinical contexts [17] [18]. It is worth noting that the integration of inference methods such as Forward Chaining and Certainty Factor, together with a knowledge base of previous cases, reflects successful approaches documented in the literature, which significantly contribute to improving the reliability of preliminary diagnoses [21] [22] [15] [16]. On the other hand, the greater dispersion observed in Prototype 4 (SD = 1.12) highlights the need to make adjustments in certain functionalities or in the interface, which aligns with the importance of proper symptom structuring and user interaction, as emphasized by [31] [32]. Nevertheless, the usability assessment and the concentration of high scores support the view that the application constitutes a practical and reliable tool, capable of optimizing early tuberculosis detection, improving patient guidance, and strengthening clinical support even before laboratory confirmation, consistent with the findings of previous expert systems [23] [24] [25] [26] [27]. Taken together, these results confirm that the system developed not only follows best practices identified in prior studies but also contributes a specialized, expert-validated approach for the early identification of tuberculosis, consolidating itself as a robust technological tool to support medical decision-making and improve timely care in settings with limited access to healthcare services.

VII. Conclusion

Tuberculosis continues to be a disease of great relevance to public health due to its high morbidity and the significant impact it generates on the population, especially in contexts with limited access to medical services. In this regard, the situation highlights the need to implement effective strategies for early detection and clinical management. Likewise, it is essential to have technological tools that facilitate medical decision-making, optimize diagnoses, and help reduce the spread of the disease in vulnerable communities. Therefore, the implementation of a web-based expert system is presented as a particularly pertinent and effective alternative. For this research, the Buchanan methodology proved to be an optimal choice, since its approach allows the entire expert system development process to be addressed in a structured manner, from the conceptualization of the problem posed to obtaining results based on the system's evaluation.

The Buchanan methodology consists of five clearly defined stages. In the first stage, identification, the problem analysis is carried out, based on the project's needs. Subsequently, the conceptualization phase focuses on outlining the internal functioning of the system, that is, how knowledge is managed and the rules it follows to generate predictions. During the

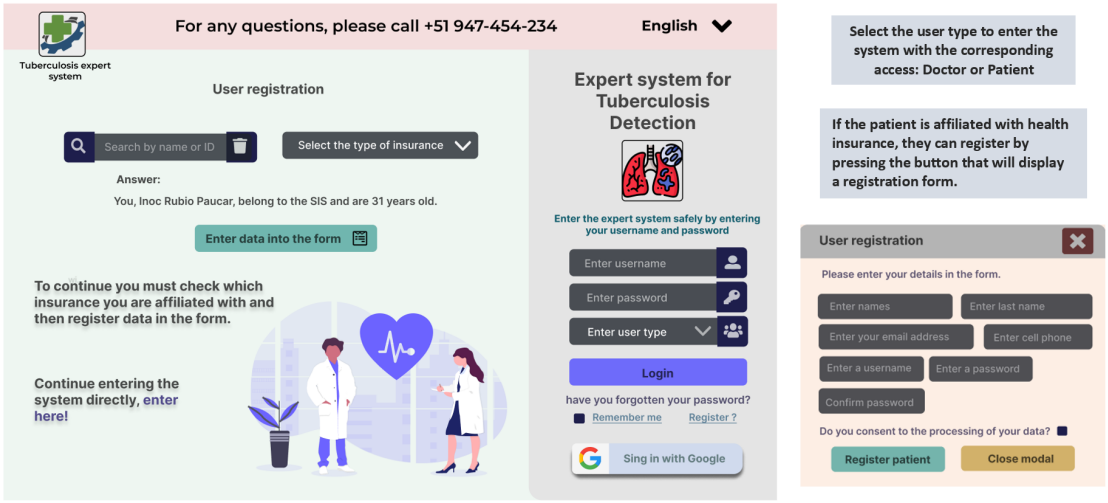
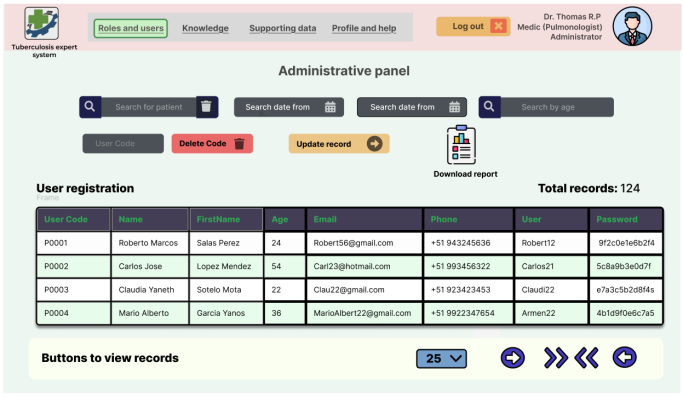
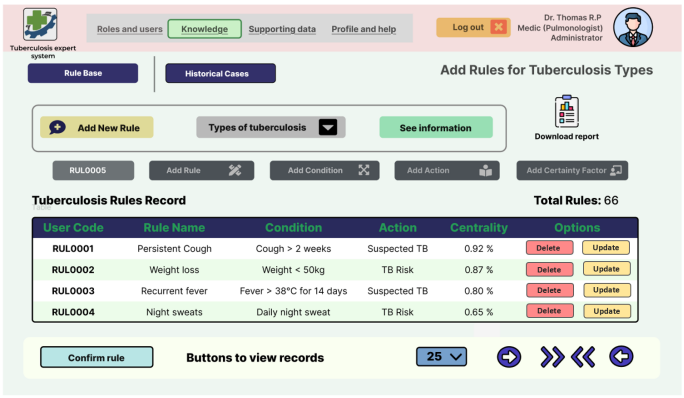


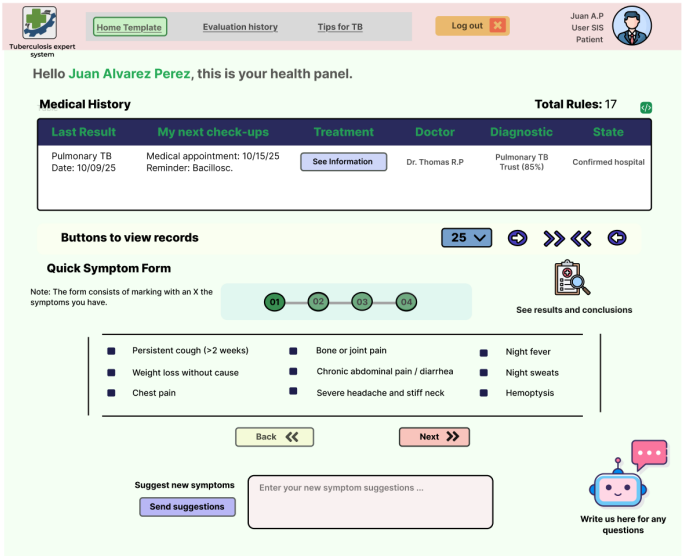
Fig. 7. Patient login and registration prototype.



(a) Administrative panel for user management.



(b) Managing rules to detect the type of TB.



(c) User panel to predict the type of TB.

Fig. 8. Prototypes of an expert web system for TB detection.

TABLE IV. EXPERT JUDGMENT QUESTIONNAIRE FOR PHYSICIANS

Dimension	Question
Usability	1. Do you consider the system interface clear and easy to use for a healthcare professional?
	2. Does the layout of menus, forms, and results facilitate quick access to clinical information?
	3. Does the system offer an intuitive design that minimizes the need for prior training?
Technical Functioning	4. Does the system respond quickly and remain stable during use in clinical practice?
	5. Do you consider the credential validation (user, role, access) secure and reliable?
	6. Does the system present technical errors or inconsistencies that may affect the user experience?
Knowledge Base	7. Is the information on symptoms, risk factors, and diagnostic criteria for tuberculosis accurate and scientifically valid?
	8. Do you consider that the rules implemented in the system adequately reflect current clinical practice?
	9. Should the knowledge base be updated more frequently to remain aligned with new clinical guidelines?
Inference Engine	10. Does the system provide diagnostic recommendations consistent with the usual clinical logic for tuberculosis detection?
	11. Do you consider the accuracy of the inferences adequate to support medical decision-making?
	12. Does the system avoid contradictions or ambiguities in the suggested results?
Expected Impact	13. Does the expert system have the potential to reduce diagnostic errors in the identification of tuberculosis cases?
	14. Do you consider it a useful tool in primary care settings or in resource-limited contexts?
	15. Do you think the expert system will contribute to improving efficiency and quality in the early diagnosis of tuberculosis?

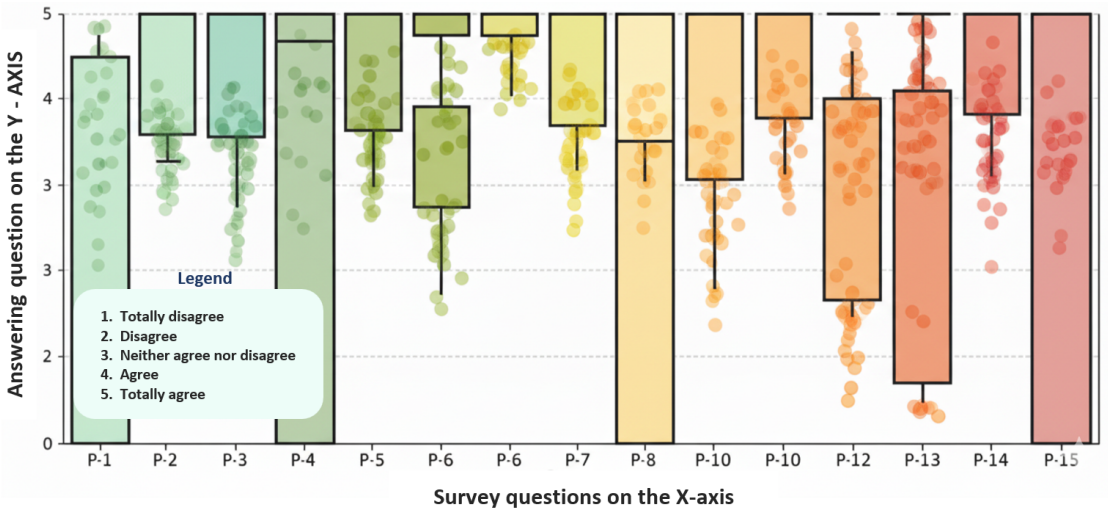


Fig. 9. Patient login and registration prototype.

TABLE V. DESCRIPTIVE METRICS OF PERCEPTION SCORES BY PROTOTYPE

Prototype	N	Mean	Standard Deviation (SD)	Median	Minimum	Maximum
P-1	20	4.35	0.93	4.5	2	5
P-2	20	4.15	0.93	4.0	1	5
P-3	20	4.35	0.99	5.0	1	5
P-4	20	4.00	1.12	4.5	2	5
P-5	20	4.60	0.68	5.0	3	5
P-6	20	4.45	0.94	5.0	2	5
P-7	20	4.65	0.75	5.0	2	5
P-8	20	4.60	0.68	5.0	3	5
P-9	20	4.60	0.68	5.0	3	5
P-10	20	4.40	0.82	4.5	3	5
P-11	20	4.65	0.67	5.0	2	5
P-12	20	4.60	0.68	5.0	3	5
P-13	20	4.80	0.61	5.0	2	5
P-14	20	4.70	0.73	5.0	3	5
P-15	20	4.55	0.69	5.0	3	5

formalization phase, the rules to be used as filters in the patient's test are precisely established, incorporating the use of the Certainty Factor to quantify the degree of confidence in the expert system's conclusions, as well as the data flow in the knowledge base. The implementation phase focuses on the development of the graphical user interface, reflecting and consolidating all the previous processes into a functional web-based expert system. Finally, the validation phase was carried out through expert judgment, using a 15-question questionnaire designed to assess the effectiveness, reliability, and usability of the system. The results showed a high overall positive consensus, with question 13 standing out for obtaining the highest mean (4.80) and the lowest dispersion (SD = 0.61), reflecting the most favorable perception and the greatest agreement among specialists. In contrast, question 4 recorded the lowest mean (4.00) and the highest dispersion (SD = 1.12), indicating aspects of the system that generated more divided opinions. Taken together, these results confirm that the web-based expert system constitutes a reliable and relevant tool to support the diagnosis of tuberculosis, backed by the direct evaluation of healthcare professionals.

Finally, understanding the complexity of early tuberculosis diagnosis is essential to strengthening healthcare delivery and improving clinical decision-making. Based on the findings obtained through the evaluation of the web-based expert system and expert judgment, it is suggested to develop complementary technological tools, such as the integration of ML algorithms, radiological image analysis, mobile applications for symptom tracking, and early warning platforms that can anticipate tuberculosis cases and facilitate timely interventions. Likewise, future studies could incorporate additional variables, explore other inference methods, and expand the knowledge base in order to optimize diagnostic accuracy and contribute to the design of more robust, accessible, and efficient clinical support systems in resource-limited settings.

ACKNOWLEDGMENT

Thanks to the University of Sciences and Humanities, which made this project a reality.

REFERENCES

[1] B. Seabra and R. Duarte, "Tuberculosis national registries and data on diagnosis delay – is there room for improvement?" *Pulmonology*, vol. 30, no. 2, pp. 130–136, 2024, doi:10.1016/j.pulmoe.2021.05.004.

[2] P. A. H. Organization, "Tuberculosis in the americas. regional report 2021," 2021, doi:10.37774/9789275126493.

[3] M. M. van der Zalm, V. W. Jongen, R. Swanepoel, K. Zimri, B. Allwood, M. Palmer, R. Dunbar, P. Goussard, H. S. Schaaf, A. C. Hesseling, and J. A. Seddon, "Impaired lung function in adolescents with pulmonary tuberculosis during treatment and following treatment completion," *eClinicalMedicine*, vol. 67, p. 102406, 2024, doi: 10.1016/j.eclinm.2023.102406.

[4] Y. Liu, Y. H. Wu, S. C. Zhang, L. Liu, M. Wu, and M. M. Cheng, "Revisiting computer-aided tuberculosis diagnosis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2316–2332, 2024, doi:10.1109/TPAMI.2023.3330825.

[5] S. Swain, A. Kumar, V. K. Vishwakarma, A. A. K, A. Mittal, and N. Wig, "Diagnosis and management of latent tuberculosis infection: Updates," *Infectious Disorders - Drug Targets*, vol. 24, no. 4, pp. 12–19, 2024, doi: 10.2174/0118715265275319231124053615.

[6] P. F. Pinto, C. S. Teixeira, M. Y. Ichihara, D. Rasella, J. S. Nery, S. O. Sena, E. B. Brickley, M. L. Barreto, M. N. Sanchez, and J. M. Pescarini, "Incidence and risk factors of tuberculosis among 420 854 household contacts of patients with tuberculosis in the 100 million brazilian cohort (2004–18): a cohort study," *The Lancet Infectious Diseases*, vol. 24, no. 1, pp. 26–46, 2024, doi:10.1016/S1473-3099(23)00371-7.

[7] W. Bai and E. K. Ameyaw, "Global, regional and national trends in tuberculosis incidence and main risk factors: a study using data from 2000 to 2021," *BMC Public Health*, vol. 24, no. 12, pp. 1–14, 2024, doi: 10.1186/s12889-023-17495-6.

[8] M. Furqan, A. H. Hasugian, and T. Elisa, "Expert system for diagnosing respiratory diseases in humans using forward chaining method," *Journal of Information System Application (Computer & Management)*, vol. 4, no. 1, pp. 80–89, 2023, doi:10.1234/jpsi.2023.0401.

[9] A. Wijayanti, F. N. Arifah, D. E. Putri, M. D. Satriyanto, and S. Sallu, "Expert system for diagnosing tuberculosis implementing case-based reasoning method," *Journal of Computer System and Informatics (JoSYC)*, vol. 4, no. 3, pp. 570–577, 2023, doi:10.47065/josyc.v4i3.3409.

[10] A. Sharma, E. Machado, K. V. B. Lima, P. N. Suffys, and E. C. Conceição, "Tuberculosis drug resistance profiling based on machine learning: A literature review," *Brazilian Journal of Infectious Diseases*, vol. 26, no. 1, p. 102332, 2022, doi:10.1016/j.bjid.2022.102332.

[11] N. Tang, M. Yuan, Z. Chen, J. Ma, R. Sun, Y. Yang, Q. He, X. Guo, S. Hu, and J. Zhou, "Machine learning prediction model of tuberculosis incidence based on meteorological factors and air pollutants," *International Journal of Environmental Research and Public Health*, vol. 20, no. 5, p. 3910, 2023, doi:10.3390/ijerph20053910.

[12] K. Godiyal and S. Pokhriyal, "Tuberculosis prediction by machine learning techniques," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 10, no. 1, pp. 92–98, 2022, doi:10.17762/ijritcc.v10i1s.5797.

[13] Z. Indra and M. R. P. Harahap, "Development of an expert system for diagnosing tuberculosis during the pandemic using forward chaining at medan johor health center," *Karismatika*, vol. 8, no. 3, 2022, doi:10.1234/karismatika.2022.0803.

[14] M. Singh, G. V. Pujar, S. A. Kumar, M. Bhagyalalitha, H. S. Akshatha, B. Abuhaija, A. R. Alsoud, L. Abualigah, N. M. Beeraka, and A. H. Gandomi, "Evolution of machine learning in tuberculosis diagnosis: A review of deep learning-based medical applications," *Electronics (Switzerland)*, vol. 11, no. 17, 2022, doi:10.3390/electronics11172634.

[15] G. W. N. Wibowo, S. Widiastuti, M. Muratno, E. Lolang, and S. Soraya, "Application of bayes' theorem method in diagnosing tuberculosis," *Building of Informatics, Technology and Science (BITS)*, vol. 4, no. 4, pp. 1782–1788, 2023, doi:10.47065/bits.v4i4.3035.

[16] J. Sapdana and Y. Henryanto, "Implementation of expert system for diagnosing tuberculosis using dempster-shafer method," *International Journal of Software Engineering and Computer Science (IJSECS)*, vol. 2, no. 1, pp. 26–32, 2022, doi:10.35870/ijsecs.v2i1.763.

[17] L. Andrade-Arenas, I. M. R. Paucar, and C. Yactayo-Arias, "Expert systems and epidemiological surveillance for tuberculosis: Innovative tools for disease prevention and control," *International Journal of Engineering Trends and Technology*, vol. 72, no. 3, pp. 72–90, 2024, doi:10.14445/22315381/IJETT-V72I3P108.

[18] Y. Luo, Y. Xue, W. Liu, H. Song, Y. Huang, G. Tang, F. Wang, Q. Wang, Y. Cai, and Z. Sun, "Development of diagnostic algorithm using machine learning for distinguishing between active tuberculosis and latent tuberculosis infection," *BMC Infectious Diseases*, vol. 22, no. 1, 2022, doi:10.1186/s12879-022-07954-7.

[19] M. Palmer, J. A. Seddon, M. M. van der Zalm, A. C. Hesseling, P. Goussard, H. S. Schaaf, J. Morrison, B. van Ginneken, J. Melendez, E. Walters, and K. Murphy, "Optimising computer-aided detection to identify intra-thoracic tuberculosis on chest x-ray in south african children," *PLOS Global Public Health*, vol. 3, no. 5, pp. 1–15, 2023, doi:10.1371/journal.pgph.0001799.

[20] R. Mehrrotraa, M. A. Ansari, R. Agrawal, P. Tripathi, M. B. B. Heyat, M. Al-Sarem, A. Y. M. Muaad, W. A. E. Nagmeldin, A. Abdelmaboud, and F. Saeed, "Ensembling of efficient deep convolutional networks and machine learning algorithms for resource effective detection of tuberculosis using thoracic (chest) radiography," *IEEE Access*, vol. 10, 2022, doi:10.1109/ACCESS.2022.3194152.

[21] N. D. Wirasbawa, C. T. P. Widjaja, C. I. Wenji, and S. Hansun, "Expert api for early detection of tb disease with forward chaining and certainty factor algorithms," *Informatica (Slovenia)*, vol. 46, no. 6, pp. 117–124, 2022, doi:10.31449/inf.v46i6.3947.

[22] X. Hu, J. Wang, Y. Ju, X. Zhang, W. Qimanguli, C. Li, L. Yue, B. Tuohetaerbaike, Y. Li, H. Wen, W. Zhang, C. Chen, Y. Yang, J. Wang, and F. Chen, "Combining metabolome and clinical indicators with machine learning provides some promising diagnostic markers to precisely detect smear-positive/negative pulmonary tuberculosis," *BMC Infectious Diseases*, vol. 22, no. 1, 2022, doi:10.1186/s12879-022-07694-8.

[23] M. R. Syahwana and R. M. Simanjorang, "Analysis of expert system using bayes method for diagnosing tuberculosis," *Jurnal Sistem Informasi, Teknik Informatika dan Teknologi Pendidikan*, vol. 1, no. 2, pp. 57–66, 2022, doi:10.55338/justikpen.v1i2.7.

[24] L. Wang, X. D. Zhang, J. W. Tang, Z. W. Ma, M. Usman, Q. H. Liu, C. Y. Wu, F. Li, Z. B. Zhu, and B. Gu, "Machine learning analysis of sers fingerprinting for the rapid determination of mycobacterium tuberculosis infection and drug resistance," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 5364–5377, 2022, doi:10.1016/j.csbj.2022.09.031.

[25] T. R. Lane, F. Urbina, L. Rank, J. Gerlach, O. Riabova, A. Lepioshkin, E. Kazakova, A. Vocat, V. Tkachenko, S. Cole, V. Makarov, and S. Ekins, "Machine learning models for mycobacterium tuberculosis in vitro activity: Prediction and target visualization," *Molecular Pharmaceutics*, vol. 19, no. 2, pp. 674–689, 2022, doi:10.1021/acs.molpharmaceut.1c00791.

[26] L. Verboven, T. Calders, S. Callens, J. Black, G. Maartens, K. E. Dooley, S. Potgieter, R. M. Warren, K. Laukens, and A. V. Rie, "A treatment recommender clinical decision support system for personalized medicine: Method development and proof-of-concept for drug resistant tuberculosis," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 56, 2022, doi:10.1186/s12911-022-01790-0.

[27] O. Hrizi, K. Gasmi, I. B. Ltaifa, H. Alshammari, H. Karamti, M. Krichen, L. B. Ammar, and M. A. Mahmood, "Tuberculosis disease diagnosis based on an optimized machine learning model," *Journal of Healthcare Engineering*, vol. 2022, p. 8950243, 2022, doi:10.1155/2022/8950243.

[28] B. M. B. Sembiring and P. M. Hasugian, "Tuberculosis disease diagnosis expert system using certainty factor method," *Journal Of Computer Networks, Architecture and High Performance Computing*, vol. 2, no. 1, pp. 129–134, 2020, doi:10.47709/cnapc.v2i1.373.

[29] A. Rachmad, N. Chamidah, and R. Rulaningtyas, "Mycobacterium tuberculosis identification based on color feature extraction using expert system," *Annals of Biology*, vol. 36, no. 2, pp. 196–202, 2020, doi:10.5555/20203239846.

[30] F. Yang, H. Yu, K. Kantipudi, M. Karki, Y. M. Kassim, A. Rosenthal, D. E. Hurt, Z. Yaniv, and S. Jaeger, "Differentiating between drug-sensitive and drug-resistant tuberculosis with machine learning for clinical and radiological features," *Quantitative Imaging in Medicine and Surgery*, vol. 12, no. 1, pp. 675–687, 2022, doi:10.21037/qims-21-290.

[31] N. Sari, V. Sihombing, and D. Irmayani, "Implementation of the bayes method for diagnosing tuberculosis," *SinkrOn*, vol. 5, no. 2, pp. 325–331, 2021, doi:10.33395/sinkron.v5i2.10903.

[32] S. P. Althomsons, K. Winglee, C. M. Heilig, S. Talarico, B. Silk, J. Wortham, A. N. Hill, and T. R. Navin, "Using machine learning techniques and national tuberculosis surveillance data to predict excess growth in genotyped tuberculosis clusters," *American Journal of Epidemiology*, vol. 191, no. 11, pp. 1936–1943, 2022, doi:10.1093/aje/kwac117.

[33] F. Weng, Y. Meng, F. Lu, Y. Wang, W. Wang, L. Xu, D. Cheng, and J. Zhu, "Differentiation of intestinal tuberculosis and crohn's disease through an explainable machine learning method," *Scientific Reports*, vol. 12, no. 1, p. 5551, 2022, doi:10.1038/s41598-022-05571-7.

[34] W. Deelder, G. Napier, S. Campino, L. Palla, J. Phelan, and T. G. Clark, "A modified decision tree approach to improve the prediction and mutation discovery for drug resistance in mycobacterium tuberculosis," *BMC Genomics*, vol. 23, no. 1, p. 46, 2022, doi:10.1186/s12864-022-08291-4.

[35] T. S. Brown, L. Tang, S. V. Omar, L. Joseph, G. Meintjes, G. Maartens, S. Wasserman, N. S. Shah, M. R. Farhat, N. R. Gandhi, N. Ismail, J. C. Brust, and B. Mathema, "Genotype–phenotype characterization of serial mycobacterium tuberculosis isolates in bedaquiline-resistant tuberculosis," *Clinical Infectious Diseases*, vol. 78, no. 2, p. 269–276, 2024, doi:10.1093/cid/ciad596.

[36] N. J. Day, P. Santucci, and M. G. Gutierrez, "Host cell environments and antibiotic efficacy in tuberculosis," *Trends in Microbiology*, vol. 32, no. 3, p. 1862–1870, 2024, doi:10.1016/j.tim.2023.08.009.

[37] J. H. Hyun, M. Lee, I. Jung, E. Kim, S. M. Hahn, Y. R. Kim, S. Lim, K. Ihn, M. Y. Kim, J. G. Ahn, J. S. Yeom, S. J. Jeong, and J. M. Kang, "Changes in tuberculosis risk after transplantation in the setting of decreased community tuberculosis incidence: a national population-based study, 2008–2020," *Annals of Clinical Microbiology and Antimicrobials*, vol. 23, no. 1, 2024, doi:10.1186/s12941-023-00661-4.

[38] P. C. Renjifo Ramos, "Situación epidemiológica de la tuberculosis en el perú," 2022, accessed: 2025-09-25. [Online]. Available: https://es.scribd.com/document/602251896/03

[39] M. M. Kumbure, A. Tarkiainen, J. Stoklasa, P. Luukka, and A. Jantunen, "Causal maps in the analysis and unsupervised assessment of the development of expert knowledge: Quantification of the learning effects for knowledge management purposes[formula presented]," *Expert Systems with Applications*, vol. 236, 2024, doi:10.1016/j.eswa.2023.121232.

[40] L. P. Wanti, O. Somantri, N. W. A. Prasetya, and L. Puspitasari, "Fuzzy expert system design for detecting stunting," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 34, no. 1, pp. 556–564, 2024, doi:10.11591/ijeecs.v34.i1.pp556-564.

[41] N. Marvi, S. Mollazadeh, F. E. Arghavanian, A. Atashi, and T. Khadivzadeh, "Designing, validation and evaluation of the expert system of "healthy menopause" and assessing its effect on the management of menopause symptoms: an exploratory mixed method study protocol," *Reproductive Health*, vol. 21, no. 1, p. 9, 2024, doi:10.1186/s12978-024-01740-1.

[42] D. P. Maison, "Tuberculosis pathophysiology and anti-vegf intervention," *Journal of Clinical Tuberculosis and Other Mycobacterial Diseases*, vol. 27, 2022, doi:10.1016/j.jctube.2022.100300.

[43] K. C. Rahlwes, B. R. Dias, P. C. Campos, S. Alvarez-Arguedas, and M. U. Shiloh, "Pathogenicity and virulence of mycobacterium tuberculosis," *Virulence*, vol. 14, no. 1, p. 1, 2023, doi:10.1080/21505594.2022.2150449.

[44] B. Zacharia and A. Roy, "Multi drug resistant tuberculosis," *European Journal of Clinical and Experimental Medicine*, vol. 17, no. 3, pp. 316–353, 2019, doi:10.15584/EJCEM.2019.3.11.

[45] J. Meghji, O. M. Kon, and A. Ainley, "Clinical tuberculosis," *Medicine (United Kingdom)*, vol. 51, no. 11, pp. 768–773, 2023, doi:10.1016/j.mpmed.2023.08.012.

[46] R. D. Kanabalan, L. J. Lee, T. Y. Lee, P. P. Chong, L. Hassan, R. Ismail, and V. K. Chin, "Human tuberculosis and mycobacterium tuberculosis complex: A review on genetic diversity, pathogenesis and omics approaches in host biomarkers discovery," *Microbiological Research*, vol. 246, 2021, doi:10.1016/j.micres.2020.126674.

[47] C. Lange, "Management of drug-resistant tuberculosis," *Deutsche Medizinische Wochenschrift*, vol. 148, no. 19, pp. 1236–1241, 2023, doi:10.1055/a-1939-0000.

[48] J. Espinosa-Pereiro, A. Sánchez-Montalvá, M. L. Aznar, and M. Espiau, "Mdr tuberculosis treatment," *Medicina (Lithuania)*, vol. 58, no. 2, p. 188, 2022, doi:10.3390/medicina58020188.

Feature-Optimized Machine Learning for High-Accuracy Ammunition Detection in X-Ray Security Screening

Osama Dorgham[1], Nijad Al-Najdawi[2], Mohammad H. Ryalat[3], Sara Tedmori[4], Sanad Aburass[5]
Prince Abdullah bin Ghazi Faculty of Information and Communication Technology,
Al-Balqa Applied University, Al-Salt, 19117, Jordan[1,3]
Department of Computer Science-College of Computer and Information Sciences,
Prince Sultan University, Riyadh, Kingdom of Saudi Arabia[2]
King Hussein School of Computing Sciences, Princess Sumaya University for Technology, 11941, Amman, Jordan[4]
Department of Computer Science, Luther College, Decorah, Iowa, USA[5]

***Abstract*—This paper introduces a machine learning system that is feature-optimized to enhance the detection of concealed ammunition in X-Ray security imaging. The system integrates advanced image analysis techniques with a cascade-AdaBoost classifier and Multi-scale Block Local Binary Pattern (MB-LBP) features, which are particularly effective for object recognition and classification in complex, high-dimensional data. The combination of these algorithms ensures robust performance in identifying ammunition types even under challenging conditions, such as variations in image quality or object orientation. The system is specifically designed for the accurate identification of various types of ammunition, including 9 mm bullets for handguns, AK-47 machine gun bullets, and 12-gauge shotgun cartridges. To support the development and testing of this system, a new dataset comprising 1,732 X-Ray images of passenger luggage was collected. This dataset is made publicly available to facilitate further research and improvement in this critical area of security technology. Experimental results demonstrate that the system achieves a high level of detection accuracy, with the ability to identify 12-gauge shotgun shells concealed in baggage with a 92% success rate. Beyond its technical achievements, this system significantly enhances the efficiency and reliability of security checks, improving the overall effectiveness of ammunition detection in real-world scenarios.**

***Keywords*—Feature optimization; ammunition detection; X-Ray images; machine learning; security imaging**

I. Introduction

In the 21st century, there are rapidly increasing needs in many urban-related areas due to the current growth of urban populations, including environmental governance, public safety, city planning, industry facilitation, resource utilization, energy conservation, traffic control, telemedicine, homecare, and interpersonal relationships, social interactions, entertainment, and communication. Failure to satisfy any of the criteria could threaten a city's ability to grow sustainably. Smart cities play a critical role in the advancement of humanity, which often depends on widely dispersed smart devices to monitor the urban environment in real-time, react quickly, establish automated control, gather data for intelligent decision-making, enable various services, and enhance the quality of urban living [1].

These days, fast, smooth, and efficient transport of people and their luggage has become very important in this fast-paced modern lifestyle. Much attention has also been paid recently to the security and safety of passengers by the aviation industry. Occasionally, national safety, economic structure, and infrastructure rely on a secure and safe airport system. Therefore, much time and effort are continuously being exerted in order to develop a safer policy and improve airport security. However, despite all these precautions, there are many terrible accidents that occur regularly indicating limitations and drawbacks in the currently used security systems [2].

In airports, security checkpoints use metal detectors or X-ray detection systems for discovering metallic objects along with high-tech plastics or ceramics, composite weapons, illegal drugs, explosives, ammunition, or other items that can be smuggled [3]. There are many places such as airports, public places, and sensitive buildings that use manual screening for detecting unauthorized and concealed items and weapons [4]. These systems can be very effective, quick, and dynamic, regardless of the enormous number of individuals accessing these areas. Dillingham [5] determined several constraints and challenges in the capability of the screening procedures to discover suspicious items on the body or the luggage of the passengers. Consequently, it can be concluded that the conventional manual screening processes that are still used at many airports are incapable of detecting all the concealed and threatening weapons because of the restrictions in the technology and in the performance of the personnel [6].

The baggage screening and inspection process takes a lot of effort and time. On the other hand, many passengers also complain that the screening procedures invade their privacy by scanning and checking their personal stuff and belongings [7]. Conventional security systems count totally on the performance, perceptions, judgment, and ability of the security personnel in making decisions [8]. However, it is argued that human recognition systems are more accurate and efficient than automated systems. Nevertheless, there are many additional factors, such as placing objects at unexpected angles, that prevent object discovery by using their images [9]. In the same context, the Congressional Research Service (CRS) report to congress has stated that several factors could contribute to limiting human performance such as tiredness,

a lack of job satisfaction or motivation, inadequate training, and unsuitable workplace conditions, which affect their performance and even their standards of success [6]. Other studies [10][11] have stated that after working for long time periods, the human eyes get exhausted and tired. In addition, when humans have been working on the same mission for long time periods, they can get bored due to monotonous tasks and lose their interest and energy. This could lead to an erroneous reading and inexpedient analysis of the information; this would result in negligence in discovering the presence of concealed weapons or such items [2]. Moreover, a cluttered X-ray image of the passenger's belongings contains many shapes and colors [12], which could overload the security personnel and make the detection of weapons or explosive devices more difficult amongst the other personal items carried by the passenger.

Metal detection machines used for detecting metallic objects at airport security checkpoints raise an alarm for every metallic object passing through them. Even though a majority of the items are harmless, they still raise false alarms. Since many false alarms are raised every minute by these metal-detecting machines, even the security personnel find it annoying and ignore them. This practice could be responsible for many of the undetected suspicious metallic objects passing through the detectors [9]. In order to provide better security at the airports, more reliable, enhanced, and fast-screening systems are necessary for inspecting baggage at airports. This system has two objectives. First, the system must be able to analyze and study every X-ray image. Then it must contain a suspicious object detection application. A novel and automated system that carries out fast and reliable baggage monitoring must be developed. It will need less processing time and provide more accurate results. Though there are many X-ray techniques for detecting suspicious objects, there are no specific ones that focus on ammunition detection. However, there are many techniques that detect the presence of concealed weapons like knives, handguns, etc., depending on the size, shape, and the material's response to the X-ray radiation. Some of the common X-ray image-processing techniques, which are used for Concealed Weapon Detection (CWD), include image enhancement and de-noising, object segmentation, shape description, and detection and recognition of weapons [11][13][14].

In this research, for the first time, a novel dataset with 1732 X-Ray images of passenger luggage has been developed specifically for this use. The public is given access to this special dataset so that other researchers can use it. Authors propose a novel automatic system that detects different kinds of ammunition concealed in passenger baggage and carry-on items. This system can be used to enhance and support the conventional inspection techniques. Literature analysis suggested that many authors have carried out CWD techniques using image sensors, IR waves, and X-rays; however, those techniques could only detect metallic objects and were unable to detect non-metallic weapons, i.e., made of ceramics or plastics such as 12-gauge shells [7].

The content of the paper can be summarized as follows: Section II presents a literature review of detecting concealed weapons and concealed ammunition using different machine vision and image processing techniques. Section III describes the implementation of a smart system that would increase safety and security levels at airports by supporting security expert's ability of 9 mm hand-gun bullets, AK-47 machine gun bullets, and 12-gauge shotgun shells. Section IV outlines the performance of the detection evaluation with some confirmatory ranking results and analysis. Finally, Section V presents the conclusions of this paper.

II. Literature Review

In order to effectively manage data analysis, data communications, and the successful implementation of complicated policies to maintain the smooth and secure management of a smart city, the efficient use of artificial intelligence (AI) and machine learning (ML) techniques is vitally important [15]. Big data analysis and the successful application of AI and ML have increased the efficiency and scalability of smart cities [16]. In order to improve the efficiency of passenger investigation and ensure safe journeys, ML-based techniques have a significant impact on the present intelligent transportation system (ITS) [16][17].

Several researchers have utilized AI-based approaches to ensure the safety of the smart city environment [18]. Based on neural networks, the authors of [19] presented an effective crime detection system for a smart city to quickly locate and assess any criminal behavior. In the same context, the authors of [20] also suggested an ML-based architecture that might be used to anticipate incidents and provide responses before they occur. To realize future trends and the use of IoT in the intelligent transportation system, the applications employed in the IoT-based transport model are thoroughly examined in [21].

Many studies have been proposed in the literature for the purpose of detecting concealed weapons using different machine vision and image processing techniques. However, few studies have addressed the detection of concealed ammos. In the study by Kase [9], the author investigated and implemented various color-coding schemes for the detection of suspicious and threatening objects using X-ray images. The author argued that the use of these techniques would help in the detection of threatening objects and would improve the efficiency of the baggage monitoring, as well as decreasing inspection time and the possibility of fatigue-based errors.

There are several studies that have addressed the automatic detection of ammunition. These studies can be categorized according to the areas of application into six categories, which are: detection of concealed weapons such as [22], detection of unconcealed weapon [23], gun detection [24], knife detection [25], weapon detection using X-ray imaging [26], and weapon detection using infrared imaging [27]. It is worth mentioning here that some studies can be grouped in more than one of those categories.

However, the methods used for the detection of concealed ammunition can be grouped into four categories: matching based [28], classifier-based [29], multisensor-based [30], and detector-based [29]. This study proposes a classification-based approach to address concealed-munition detection using X-ray imaging.

Regarding the classifier-based approaches, weapons, ammunition, and bullets are segmented from the image, and then the region of interest (ROI) are utilized to extract features that

will be used later for classification [31]. The main role of the classification approach is to classify images as positive or negative. The presence of the ammunition in the input image indicates positive and the absence of the ammunition means negative. Thus, segmentation of ROI is considered a critical phase in the classifier-based approaches to extract the required features for classification.

Several studies employed a classifier-based approach for the purpose of weapons detection and different traditional techniques were utilized in these studies to extract features from the segmented ROI [32]. These techniques include active contour shape, Fisher's linear discriminant, watershed level sets, Fuzzy connectedness, k-means algorithm, and region growing [33]. On the other hand, there are other studies that employed deep learning models as a base for the classification process. Lai et al [34] firstly presented in 2017 a deep learning approach in the field of gun detection to classify images. Their approach used the GoogleNet pre-trained model to train input images. The authors of [26][35] carried out a set of pre-trained deep learning models for classification of images that include weapons and guns. Either using the traditional techniques or deep learning techniques to extract features, both of them lead to inefficient accuracy in the detection of weapons for the small size of weapons. Moreover, most of the published studies in this domain addressed the issue of weapons detection (i.e., that have a relatively large size) but neither addressed ammunition/bullets detection nor the types of ammunition.

Sadah et al. [11] proposed a hybrid technique for improving digital X-ray images by combining the frequency domain enhancement and optimal spatial image techniques. In their method, they selected these frequency domain enhancement techniques: Gaussian and Butterworth high and low pass filters, and as for the spatial domain methods: power-law transform, Histogram equalization, and Negative transform. They carried out more than 80 probable combinations, and some combinations optimally enhanced the original X-ray images. In the study by Khan and Chai, the authors proposed an approach that combined the high and the low-energy X-ray images, denoised the images, and improved the fusion image using histogram specifications, and it also improved the contrast. Their results showed that the final image contained all the minute details; it had no background noise and had enhanced contrast. These images could be easily interpreted and automatically segmented by security personnel. This will result in decreasing the false alarm rate during the X-ray baggage inspection.

Another study carried out by Al-Najdawi [2] proposed another approach that detected different types of bullets and shotgun shells. Different image processing techniques were applied in the detection process, which includes image enhancement, conversion, labeling of connected components, and geometric distance calculations. The author built a dataset comprising 150 X-Ray images of passenger's luggage that contain different types of ammunition and proposed a wideband millimeter-wave imaging system which carried out a fast inspection of the baggage for detecting concealed weapons, explosives, handguns, etc.

Image fusion techniques combine relevant information from many images, those techniques can be classified into the spatial domain and the transform domain. The spatial domain method includes the averaging approach, principal component analysis, the Brovey method, and the Hue-Intensity-Saturation (HIS) transformation approach. The high pass filtering-based processes are also included in the spatial domain approaches [36]. On the other hand, the transform domain approaches can easily handle the spatial distortion after image fusion. In the work proposed by Xue et al. [37], the authors developed an algorithm that merged the color visual images and their corresponding infrared (IR) images for CWD. This color image fusion algorithm proved to maintain the resolution of the image and could incorporate the detected weapons in the image while maintaining the natural colors of the image. In a study by Achanta et al. [38], the authors developed a method known as parametric mixing. In their study, they created a model which improved the detection of objects (i.e., metallic or non-metallic) or concealed weapons using a nonlinear acoustic technique. In their method, ultrasonic waves were used for creating a localized zone, in which the nonlinear interaction generated a low-frequency acoustic wave that could go through the clothing in a better manner when compared to direct ultrasonic waves.

Microwaves can be used for communication, i.e., for transmitting long-distance telephone signals and in satellite television. They are also used for detecting concealed metal. Elsdon et al. [39] suggested an approach for detecting the concealed metallic objects like weapons by using an indirect holographic process, which was operational at the optical frequencies. This approach could be applied to the image objects at the microwave frequencies, and it could determine the shape and location of the concealed objects. Since the microwave imaging methods is very expensive, researchers have proposed a simple and inexpensive method of using this technique for CWD.

One of the biggest problems facing researchers in the field of ammunition identification is the lack of publicly available benchmark datasets according to [30]. As a part of the contribution in this research, a novel dataset of 1732 X-Ray images of passenger luggage has been produced especially for this purpose. This unique dataset is made available to the public so that other scholars may use it.

III. Feature-Optimized Approach for Ammunition Detection

A. Feature Selection Techniques

The proposed system for the detection of ammunition from X-ray images would be as effective as the features that are extracted from the images. Due to the complexity of the images and the necessity to distinguish between the hidden ammunition, feature selection was an important step in the model training and improving its performance. In this study, the MB-LBP features were chosen based on their efficiency in describing the micro and macro-structures that are seen in the X-ray images. These features are especially useful for the identification of separate bulk density and shape of different types of ammunition. In the feature selection, the numerous possibilities of features such as texture descriptors, the shape-based features, and the intensity gradients were considered. By cross-validation, only the useful features were kept in the model. To do so, we sought to maximize the distinction

between the selected features while reducing overlapping as much as possible to include only the most important features for the classification process. This was an iterative process in which features had to be assessed not only by their ability to separate classes but also as components of a set of features that defines the model. The purpose was to achieve high accuracy to guarantee that the final set of the features is both as small as possible and as accurate as possible.

The selected features were also chosen for their insensitivity to variations in conditions typical for real-world security scenarios, light conditions, object orientation, occlusions in X-ray images, etc. Thus, it was ensured that the set of features being used was rather stable with regards to these changes and by using this set of features, the system was made more robust to changes in the environment and situations. However, the focus was also paid to the interpretability of the chosen characteristics that were used for the definition of the described model. This was important so that actions of the system could be explained and be accepted by security personnel. For example, the features of the MB-LBP include clear and interpretable representations of the visual data on which the system makes its classification, which facilitates an understanding of the process that led to the classification. This transparency is important for practical implementation in serious security scenarios where not only the decision but the rationale for the decision may be critical for use. This concentration on feature selection not only enhanced the model's accuracy but also developed confidence that the system could work in real-time, making it an important tool in supporting security in areas like airports and essential installations.

B. Ammunition Detection: 9 mm Handgun and AK-47 Bullets

The greyscale images are produced by measuring the light intensities at every pixel in the band of an electromagnetic spectrum (e.g., IR, X-ray, visible light, or UV). Under these circumstances, they are monochromatic only when a light frequency is captured; however, they can also be synthesized using colored images. Grey X-rays images are an old and popular method for carrying out security checks. During baggage inspection, the luggage is exposed to small ionizing radiation doses, which results in images of the luggage's interiors [40]. However, digital X-ray images can be degraded due to low contrast and blurring, which is caused by the complex items in the luggage. Also, scattering of the X-rays and electrical noises can affect the monitoring and the analysis of the images by some systems or humans. Hence, it becomes necessary to improve the X-ray image details in order to improve the visual quality of the images [41]. Ammunition is smaller in size as compared to the other personal items in the luggage; hence, the X-ray images must be improved as in the following step.

1) X-ray image enhancement: Methods of enhancing the X-ray images within a spatial domain can be divided into three groups: point processing techniques, histogram-based processing techniques, and mask processing techniques [42]. In this study, the authors have used the histogram equalization and the unsharp masking techniques. Histogram equalization is a contrast enhancement process, which adjusts the pixel intensities for generating a new and enhanced image that shows a better local contrast. On the other hand, the unsharp masking technique sharpens the images by subtracting the smoothed

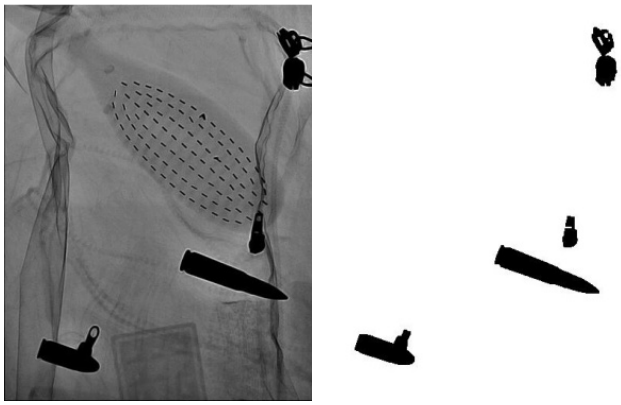


Fig. 1. Samples of the X-ray images of: (a) 9 mm handgun bullets and AK-47 machine gun bullets; (b) Resultant binary image.

image from the original image. A histogram can be created for an image based on the probability of the occurrence of the pixel intensity as described earlier [43]:

$$P_r(r_k) = \frac{n_k}{MN} \tag{1}$$

Where $P_r(r_k)$ refers to the probability of the occurrence of the intensity level r_k within a digital image, MN refers to the total pixels within an image, and n_k represents the number of the pixels with an intensity r_k. New intensity values can be estimated for every intensity level using the transformation as follows:

$$s_k = T(r_k) = (L-1)\sum_{j=0}^{k} p_r(r_j) \tag{2}$$

Where s_k refers to the new intensity level. It is estimated for every pixel in an image. In addition, the processed image can be obtained after mapping every pixel in an input image (having intensity r_k) into the corresponding pixels (with intensity level s_k) into the output images [23]. The unsharp masking procedure is carried out after blurring the original images and then creating the mask after subtracting the blurred images from the primary images as follows:

$$W_{mask}(x,y) = f(x,y) - f'(x,y) \tag{3}$$

Where $W_{mask}(x,y)$ is the mask, $f(x,y)$ represents the primary image, and $f'(x,y)$ refers to the blurred image. Thereafter, the mask is added to the primary image as follows:

$$W(x,y) = f(x,y) + W_{mask}(x,y) \tag{4}$$

Where $W(x,y)$ refers to the resultant unsharp image.

2) Ammunition detection: Ammunition detection is initiated by converting enhanced images into binary images with the help of the Otsu threshold process [43]. The images are denoised using some morphological operators. Then, the objects connected at the borders are eliminated from the resultant binary images. Finally, certain holes in the images are removed by morphological operators [43]. The Otsu threshold process searches for the threshold that decreases the intra-class variance, which can be defined as the weighted sum of the variances of the two classes:

$$\sigma_w^2(t) = w_1(t)\sigma_1^2(t) + w_2(t)\sigma_2^2(t) \tag{5}$$

Where the weights w_i refer to the probabilities of the classes, which are divided by the threshold t, and σ_i^2 refer to the variances of the class. Fig. 1 describes the final binary images after techniques like histogram equalization, unsharp masking, and the Otsu's thresholding method were applied to the X-ray images.

For the ammunition detection and classification, the connected regions in the binary images can be determined with the help of the connected-component labeling algorithm [44]. Here, the authors have applied the 8-connected component labeling algorithm. In the case of the binary image, the connected pixel areas can be classified after the connected component labeling operator has scanned the image in every row till it reaches a point p (where p represents the pixel labeled at any step during scanning). After recognizing this condition, the algorithm investigates the 4 neighbors of the point p, which were already scanned. The point p is labeled based on the collected data as shown below [29]:

- If all the neighboring points are 0, a new label is allocated to p.
- An exclusive label is allocated to every class with the help of the Floyd-Warshall algorithm.
- The image is scanned again, and every label gets replaced by the label that is allocated to the equivalence classes.

Fig. 2. Some samples of the X-ray images after ammunition detection: First row, 9 mm handgun bullets, AK-47 Machine gun bullets, and the 12-Gauge shotgun shells. Second row, resultant binary images.

In the next step, the properties of each connected component are estimated. Here, the authors have measured the properties for every labeled region of the label matrix. The positive integer elements present in the label matrix are seen to correspond to other areas. For example, the elements in the label matrix equivalent to 0 correspond to region 1, while the elements in the label matrix equivalent to 1 correspond to the region 2. Every labeled connected component gets indicated by a bounding box as described in Fig. 2. In order to determine if the bounding box contained ammunition images, a classification system was introduced. In this study, for testing purposes, the authors classified the ammunition in three classes: AK-47 machine gun bullets, 9-mm handgun bullets, and 12-gauge shotgun ammunition.

In the case of the AK-47 machine gun bullets and the 9-mm handgun bullets, the ammunition is classified based on their fixed bullet size ratio (i.e., length: width). In Table I, the dimensions and ratio of the bullets are summarized.

TABLE I. DIFFERENT DIMENSIONS AND RATIOS OF THE SELECTED AMMUNITION FOR TESTING OF THE DEVELOPED SYSTEM

Dimension and Ratio	9mm	AK-47	12-Gauge
Length (range)	2.781-2.969	5.600-5.620	5.1-7.6
Width	1.01	1.135	13.41
Ratio	2.75-2.93	4.93-4.95	3.80-5.66

For determining the bullet ratio in Table I, every label is estimated with the help of the Euclidean distance measure D_e for the points p and q that are located at the coordinates of (x, y) and (s, t), respectively, as:

$$D_e = \sqrt{(x-s)^2 + (y-t)^2} \tag{6}$$

Euclidean distances are used for measuring the height D_{e1} and the width D_{e2} of every object in the label, while the height:width is estimated as $R = D_{e1}/D_{e2}$. Thereafter, this ratio is compared to the results in Table I, after which the object is classified. Fig. 2 describes some samples in which the algorithm could accurately identify 9 mm handgun and AK-47 machine gun bullets, with precision 95%. Fig. 3 shows some sample images of the detected ammunition.

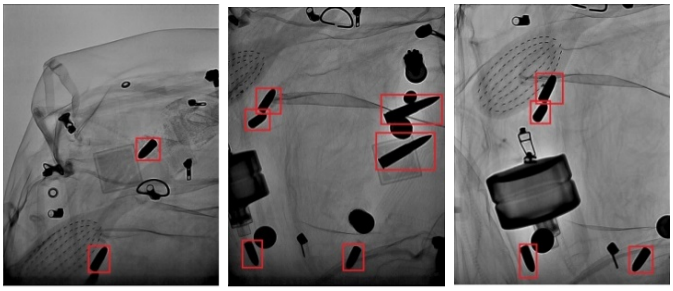


Fig. 3. Some sample images of the detected ammunition.

C. Feature-Based Identification of 12-Gauge Shotgun Shells

In this research, the authors attempt to identify the 12-gauge shells, as their structure differs from other kinds of ammunition such as the 9-mm handgun and the AK-47 machine gun bullets. The 12-gauge bullets have a plastic casing; hence, the bullets appear transparent in X-ray images. Moreover, as shown in Fig. 4, the wad creates a space between the powder

and shot which prevents its visual recognition. Therefore, any orientation change in the 12-gauge position results in different shapes, as shown in Fig. 5. On the contrary, other bullets have a solid metal body, which can be identified in the X-ray images as a solid known structure.

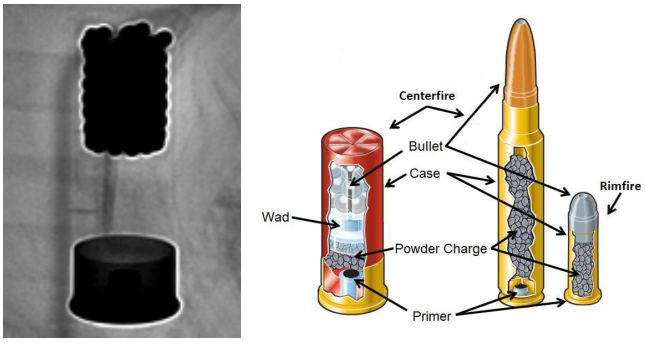


Fig. 4. (a) Structure of the 12-gauge ammunition, AK-47 machine gun bullet, and 9-mm bullet. (b) X-ray images of the 12-gauge ammunition.

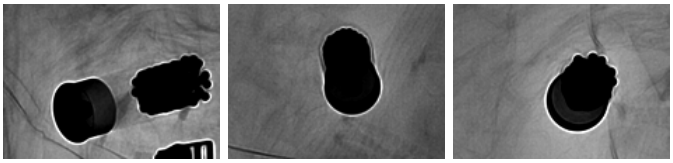


Fig. 5. Images of the 12-gauge ammunition from various angles. Image courtesy: Al-Najdawi and Tedmori.

In this study, a Cascade-AdaBoost classifier with MB-LBP features has been implemented for detecting the variable 12-gauge ammunition shapes.

AdaBoost classifier, proposed by Freund and Schapire [45][46], is a type of parallel classifier that combines various linear weak classifiers. This offline classifier self-adaptively increases the number of weak classifiers, which helps improve the overall classification accuracy. Every weak classifier is seen to focus on the 1-D classification of the input feature vectors. Moreover, the AdaBoost algorithm makes use of a minimal error for estimating the weight value of the weak classifier, then it uses the weight value of each training sample input. Then, it passes the weight value to the newly added subsequent weak classifier. Fig. 6 describes the AdaBoost classifier architecture [47].

If the input feature vector is represented by f, then the number of the weak classifiers is represented by $h_t(f)$, $t = 1, \ldots, T$, and the corresponding weight values of every weak classifier are represented by β_t, $t = 1, \ldots, T$. The final strong classifier, $H(f)$, can be expressed as follows:

$$H(F) = \text{sign}\left\{\sum_{t=1}^{T} \beta_t h_t(f)\right\} \tag{7}$$

One main limitation of the AdaBoost classifier is that it has a higher false alarm rate. Thus, Viola and Jones [48] proposed the cascade-AdaBoost classifier to decrease the false alarm rate. Fig. 7 shows the architecture of the cascade-AdaBoost classifier, where every classifier was an AdaBoost classifier.

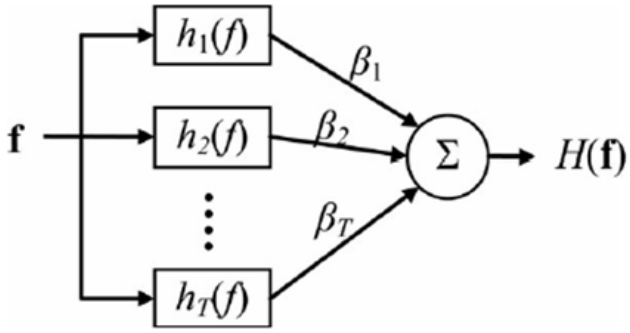


Fig. 6. Architecture of the AdaBoost classifier [47].

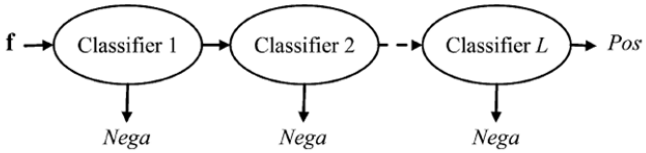


Fig. 7. Cascade-AdaBoost classifier architecture [48].

When the input feature vector is passed through the first AdaBoost classifier and is determined as the negative example, it is removed from the training set before it enters the second AdaBoost classifier. This decreases the number of the examples and increases the layers for removing the negative examples. However, when passing the input feature vector through the first AdaBoost classifier, if it is identified as a positive example, then the training example enters the second layer of the AdaBoost classifier and is classified further until it passes to the last layer. Hence, the cascade AdaBoost classifier displays a high detection rate and a lower false alarm rate since it uses many classifiers [49].

AdaBoost learning is applied for the selection of some weak classifiers and then combines them into the classifier to determine if the image contains the desired object [50]. Several algorithms can be used for feature extraction, like the Principal Component Analysis (PCA) [51], the Linear Discriminant Analysis (LDA) [52], the Independent Component Analysis (ICA) [53], and the Gabor filter [54]. However, all of these algorithms have high computational costs for convolving the object images and can be inefficient with regard to memory and time to extract features compared to the Local Binary Pattern (LBP) [55]. The LBP method was initially proposed by Ojala et al [56], and it was used for the texture description. This method can be used for many applications, such as facial detection and recognition [57][58], classification of gender and age [59][60], analyzing facial expressions [61], and estimating the head pose [62]. Some main properties of the LBP features

IV. Discussion and Experimental Results

A. Comparative Analysis of Feature Selection Techniques

The reasons for choosing the MB-LBP features in the study are based on the performance of these features in capturing the required level of details that are crucial for the identification

of ammunition in X-ray images. That is why, although there are many other techniques such as Gabor filters, Histogram of Oriented Gradients (HOG) and so on, or traditional Local Binary Patterns (LBP), in this work, the MB-LBP was chosen because of the high accuracy in combination with acceptable computational time.

Some of them, such as the Gabor filters, are particularly efficient in texture analysis and have been applied in a large variety of image processing applications. However, their computational complexity is a major disadvantage in real-time environments, especially when analyzing high-dimensional data such as X-ray images. Likewise, the HOG features are very good for edge and gradient information and are used for object detection; however, they may not be as effective in discriminating between small and closely grouped objects such as ammunition.

Although traditional LBP is faster in computation and requires fewer parameters, it focuses only on the local binary patterns without regard to the spatial relation around the center pixel, which may cause a problem in complex scenes where objects are overlapping or images have different quality. The MB-LBP approach improves upon the standard LBP approach by including larger patterns, better suited to handle the difficulties posed by X-ray imagery in security systems.

Consequently, the features of the MB-LBP were chosen for this study because of their ability to give more detailed descriptions of the X-ray images in addition to capturing more global characteristics of the images. This is in concordance with the characteristics of the ammunition detection task where precision and speed are critical. Another reason for our choice of MB-LBP was based on the observation that it has been widely used in similar applications and is generally robust for different conditions.

Comparing MB-LBP directly with other techniques was beyond the scope of this study. However, the literature shows that MB-LBP is suitable for the present task. This feature selection method is also balanced in terms of the performance and computation, and that is why its application in developing a real-time detection system in the security sector is advisable.

B. Experimental Results

In the experiments provided in this research, positive and negative samples were required in the training procedure. Positive samples (548 images that contain more than 1036 of 12-gauge ammunition) were obtained through labeling the 12-gauge ammunition from the X-ray images. Negative samples (1184 images) were extracted from X-ray images which do not contain the 12-gauge ammunition. The X-ray images of the 12-gauge ammunition that have been used in the experiments were taken from various view angles as shown previously in Fig. 5.

In the following experiments, an eighteen-layer cascade-AdaBoost classifier using MB-LBP features is obtained and then experimented over the provided dataset (i.e., offline process). Then a real-time process starts depending on the previous features which have been obtained from the MB-LBP to classify the required objects. Using this framework, it was able to process more than 27 frames per second on a machine of 8 GB RAM and 3.2 GHz processor speed. Sample of the results are shown in Fig. 8.

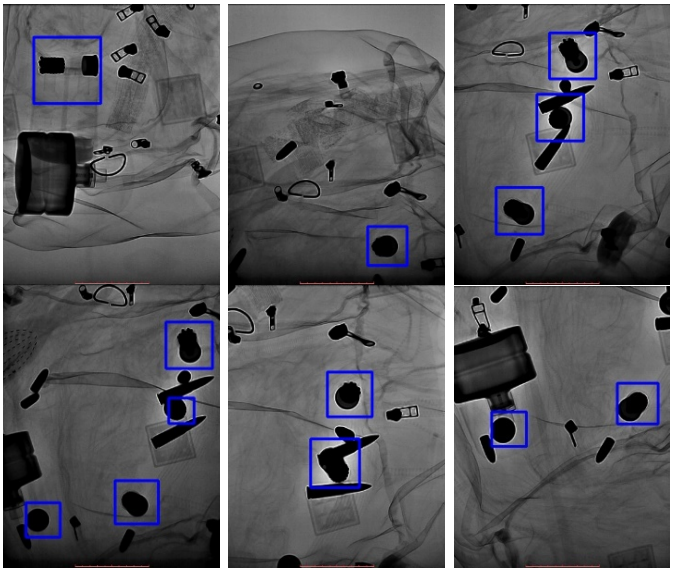


Fig. 8. Sample of successful labeling for different numbers of 12-gauge ammunition of different orientations.

However, in some cases, the proposed system failed to label the 12-gauge ammunition correctly by selecting different objects that contain similar features to the 12-gauge ammunition, as shown in Fig. 9.

In order to evaluate the performance of the proposed system, some statistical measures (i.e., accuracy, sensitivity, specificity, and Matthew's correlation coefficient) were calculated as follows:

- Accuracy has been calculated for the 12-gauge ammunition detection using the formula [63]:

$$\text{Accuracy} = \frac{\Sigma\ (TP + TN)}{\Sigma\ (TP + TN + FP + FN)} \times 100\% \tag{8}$$

  Where:
  - TP = True Positive (actual 12-gauge ammunitions that were correctly classified as 12-gauge ammunitions).
  - FP = False Positive (other objects that were incorrectly labeled as 12-gauge ammunitions).
  - FN = False Negative (12-gauge ammunitions that were incorrectly not labeled).
  - TN = True Negative (all the remaining objects, correctly classified as non-12-gauge ammunitions).

  In the proposed system, Accuracy = 92.8%, where: TP = 960, TN = 1151, FP = 88, and FN = 76.
- Sensitivity or true positive rate (TPR) relates to the ability to identify positive results using the formula [63]:

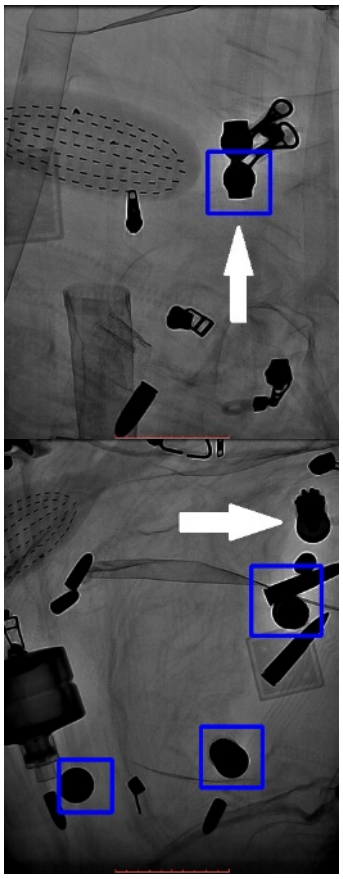


Fig. 9. Some of the cases where the system was unable to label the 12-gauge ammunition were in (a) labeling a different object or (b) missing the object completely.

$$\text{TPR} = \frac{TP}{TP + FN} \times 100\% \quad (9)$$

TPR = 92.66%, where: TP = 960 and FN = 76.

- Specificity (SPC) or true negative rate relates to the ability to identify negative results using the formula [63]:

$$\text{SPC} = \frac{TN}{TN + FP} \times 100\% \quad (10)$$

SPC = 92.89%, where: TN = 1151 and FP = 88.

- Matthew's Correlation Coefficient (MCC) was used in the machine learning for measuring the quality of the binary (2-class) classification [63]. This method considered the true and the false positives or negatives and was considered the balanced measure that could be used for the differently sized classes. MCC was used as a correlation coefficient for the observed and the predicted classification systems and generated values ranging between +1 to -1. A coefficient value of +1 showed a perfect prediction, while a value of 0 indicated a random prediction, and a value of -1 indicated a complete disagreement between the observed and predicted values. MCC is directly estimated using the confusion matrix, as follows:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (11)$$

MCC = 85.5%, where: TP = 960, TN = 1151, FP = 88, and FN = 76.

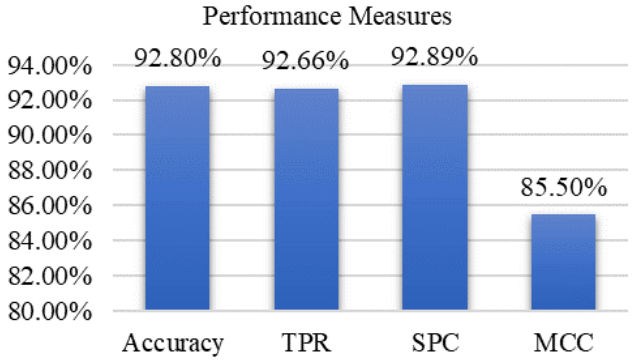


Fig. 10. Graphical representation for the performance measures (i.e., Accuracy, TPR, SPC, and MCC).

Different applications could benefit from the results of this research by applying the MB-LBP offline process as in [36][37]. Fig. 10 shows graphical representation for the performance measures (i.e., Accuracy, TPR, SPC, and MCC).

C. Comparative Analysis

In this part we compare our results with those of the related paper, "Detection of threat objects in baggage inspection with X-ray images using deep learning" by Saavedra et al. [64]. Both studies, however, address threat detection in X-ray images, and their methodologies and performances differ from each other. For this, Saavedra et al used deep learning models like YOLOv3, SSD, and RetinaNet trained on real as well as synthetic X-ray images artificially generated by PGGAN and superimposition methods. Across all four threat categories (guns, knives, razor blades, and shuriken), mAP was 80.0%. Instead, our system includes a cascade-AdaBoost classifier using the accuracy of the Multi scale Block Local Binary Pattern (MB-LBP), and is tuned for ammunition detection. The methodology described results in a 92.8% accuracy detecting 12-gauge shotgun shells, 92.66% sensitivity and 92.89% specificity.

However, guns and shuriken had high precision (96.3%, 93.7%) according to Saavedra et al. but knives with elongated shape and occlusion problems achieved low precision (76.2%). However, our system retains strong performance in detecting 12-gauge shotgun shells, which are very difficult to detect since the shells are covered in plastic casing and have multiple orientations. Cascade-AdaBoost classifier with MB-LBP features are capable of overcoming these challenges, reaching 92.8% accuracy and 85.5% MCC. In a particular dataset with similar threat objects to the training set, Saavedra et al. used their model and achieved high precision. But their model was not able to account for significantly different objects from what they trained from. Table II shows comparative analysis of

TABLE II. COMPARATIVE ANALYSIS: FEATURE-OPTIMIZED MACHINE LEARNING SYSTEM VS. DEEP LEARNING-BASED THREAT DETECTION

Aspect	Our System (Feature-Optimized ML)	Saavedra et al. (Deep Learning-Based)
Methodology	Cascade-AdaBoost classifier with Multi-scale Block Local Binary Pattern (MB-LBP) features.	Deep learning models (YOLOv3, SSD, RetinaNet) trained on real and synthetic X-ray images (PGGAN).
Dataset	1,732 X-ray images of passenger luggage with concealed ammunition (publicly available).	GDXray dataset with simulated X-ray images (real and synthetic).
Threat Objects Detected	9 mm handgun bullets, AK-47 machine gun bullets, 12-gauge shotgun shells.	Guns, knives, razor blades, shuriken (ninja stars).
Performance Metrics	Accuracy: 92.8% (12-gauge shells), Sensitivity: 92.66%, Specificity: 92.89%, and MCC: 85.5%.	mAP: 80.0%, Guns: 96.3%, Knives: 76.2%, Razor blades: 86.9%, and Shuriken: 93.7%.
Strengths	High accuracy for 12-gauge shotgun shells. And, robust to variations in orientation and occlusion.	High precision for guns and shuriken. And, scalable to multiple threat categories.
Challenges	Misclassification of objects with similar textures. And, reliance on single-view X-ray images.	Struggles with elongated objects (e.g., knives). And, limited generalization to unseen object shapes.
Generalization	Strong generalization to diverse orientations of 12-gauge shells. And, adaptable to real-world scenarios.	Limited generalization to objects with different shapes or orientations.
Real-World Applications	Suitable for airport baggage screening. And, reduces reliance on human operators.	Effective for detecting large threat objects (e.g., guns, shuriken).
Future Work	Integration of multi-view imaging. And, combining with deep learning models (e.g., YOLOv3).	Improving detection of small or occluded objects (e.g., knives).

feature-optimized machine learning system and deep learning-based threat detection.

In addition, we have shown here that our system can accurately identify 12-gauge shotgun shells; however, objects with similar appearances (in this case, plastic casings) were sometimes wrongly identified as ammunition. This is the problem of separating two closely related objects in cluttered X-ray images. Using the fixed bullet size ratio for classification, the system had 95% accuracy for 9 mm handgun bullets and AK 47 machine gun bullets. The elongated shape of knives typical of this type of objects, though, remains a challenge and future work may involve multi-view approaches to increase detection accuracy of such objects. The system is able to detect 12-gauge shotgun shells in various orientations and successfully detected 12-gauge shotgun shells in Fig. 8. Fig. 9 shows failed cases of the system in labeling of ammunition, primarily because of the presence of overlapping objects or similar texturing. They are valuable cases for advancing improvements, e.g., adding extra features or more refined segmentation. Due to the variable orientation and plastic casing of 12-gauge shells, the system does not perform as well with 12-gauge shells compared to 9-mm and AK-47 bullets, which are all about the same size and shape.

We show that our system is very robust to variations of the object orientation, occlusion, and image quality, and we also experimentally demonstrate that our system also facilitates the acquisition of stereo images of patients with imperfect eyes. In addition to capturing both micro and macro structures in an X-ray image, the MB-LBP has the capacity of reliable detection in complex scenarios. However, Saavedra et al.'s model did not perform well with objects when they were rotated or occluded, such as knives. This points out the importance of feature selection and necessity of models for dealing with diverse object orientations.

Our system works with high accuracy and robustness and thus is suited for security applications in real-world situations, e.g. for airport baggage screening. Security checks are significantly more efficient and reliable due to the increase in the ability to detect 12-gauge shotgun shells with an accuracy of 92.8%. Future work may also test our system with multi-view X-ray imaging, as suggested by Saavedra et al., to further improve these object detection accuracies when knives and shuriken are used.

One major shortcoming of our system is that it relies on single-view X-ray images, which might cause object that has complex shape or is under heavy occlusion to be missed. This limitation will be addressed in the future work through multi-view approaches. Further, objects with similar texture or shape are misclassified from time to time as ammunition. It is quite possible to mitigate this problem by incorporating advanced segmentation techniques or additional features. Applications in the future include adding multi-view X-ray imaging to provide a more global description of objects and enhancement of detection accuracy for hard cases. Further performance improvement is possible by combining our feature-optimized approach with deep learning models like YOLOv3 or RetinaNet because they are powerful models that are more efficient in classifying small or occluded objects. The generalization capability of the system could be enhanced if the dataset were expanded to include a greater variety of ammunition types and challenging scenarios.

V. CONCLUSION

Feature selection is a critical element in the improvement of the performance of machine learning systems, especially in tasks such as the detection of ammunition in X-ray security imaging systems. In this study, the choice and tuning of the features were critical in constructing a reliable system that can effectively detect concealed ammunition with a high level of accuracy. The features of MB-LBP were used and combined with the cascade-AdaBoost classifier in order to extract the most important information from the X-ray images and guarantee that the system can work in real-time environments. Safety and security of passengers is one of the important components of contemporary transport systems. Measures of security are tight, but threats and contraband are a reality in sensitive environments. X-ray machine systems at security checkpoints are very important in detecting banned items such as weapons and sharp objects. These systems may contain at least semi-automatic elements for metal detection. However, there is a significant gap in their capability of identifying concealed ammunition, which is a critical security threat. In an attempt to fill this research gap, this paper presents an automatic system designed to help security professionals detect different types of ammunition.

The main focus of the system is a profound feature selection method that enhances the accuracy and efficiency of the detection phase. The selection of MB-LBP features was especially useful in obtaining the fine detail required to differentiate between the types of ammunition with clarity, including 9 mm handgun bullets, AK-47 machine gun bullets, and 12-gauge shotgun shells. This feature optimization, in addition to improving the performance of the model, also decreases the system's computational intensity, making the system appropriate for real-time security applications. The authors have created their own dataset of 1,732 X-ray images of passengers' belongings that contain concealed ammunition. This dataset, now open to the public, promotes more research and development and brings together the security and machine learning domains. Thus, in addition to meeting the need for a reliable resource to support security specialists' work, the study helps to progress science and technology in the field of automated threat detection.

The system uses a number of sophisticated image processing techniques such as image enhancement, image conversion, labeling of components, and estimation of geometric dimensions. These techniques are very essential since they help to overcome the problem of variability of the X-ray image quality, orientation of objects, and the various complex backgrounds that may be observed in the passenger's luggage. The integration of the cascade-AdaBoost classifier with the MB-LBP features makes it possible for the system to perform effective and accurate detection of concealed ammunition. The combination of these mentioned methods leads to an overall effective, efficient, and flexible solution that can be implemented into different security scenarios.

In classification and recognition of concealed ammunition, the system performs quite well, especially for the 9 mm handgun bullets and the AK-47 machine gun bullets, with at most 92% accuracy in detection. Such a high level of accuracy shows how the system can greatly advance the detection abilities at the security checkpoints, with little reliance on human operators and the likelihood of human errors. The capacity to accurately identify ammunition and to do so without human intervention increases the likelihood of safety and security for passengers, especially within areas of vulnerability such as airports and other stations. Besides the efficiency, the system's architecture focuses on scalability and interpretability, which means that the developed system can be implemented in different real-world settings. It can be integrated into other security systems and structures, and it is relatively easy to modify the architecture as new threats appear on the scene. Furthermore, the system is very interpretable, and this means that security personnel can trust the decisions made by the system, which is very important when the system is being implemented in important settings.

This paper demonstrates that, in today's security systems, it is possible to enhance the performance of advanced feature selection and machine learning techniques. By focusing on the particular case of concealed ammunition searching, the authors have presented a useful instrument that could help to enhance public security and develop existing security technologies. Future work may look at applying such a scheme to identify other forms of prohibited items, as well as incorporating it within larger security systems as a way of expanding, generalizing, and ultimately increasing the practical applicability and effectiveness of the method.

Conflict of Interest

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 paper.

Data Availability

The X-Ray images data used to support the findings of this study have been deposited in the Automatic-Ammo-Detection- repository (https://github.com/reyalat-bau/Automatic-Ammo-Detection-).

References

[1] Y. He, I. Stojmenovic, Y. Liu, and Y. Gu, "Smart city," *Int J Distrib Sens Netw*, vol. 2014, pp. 1–2, 2014.

[2] N. Al-Najdawi, "A concealed ammo detection system for passengers luggage screening," in *2014 International Conference on Multimedia Computing and Systems (ICMCS)*. IEEE, 2014, pp. 159–164.

[3] K. Leone and R. Liu, "Validating throughput analysis of checked baggage screening performance," *Transportation Research Board for Presentation and Publication*, pp. 1–16, 2005.

[4] B. Khajone and V. Shandilya, "Concealed weapon detection using image processing," *Int. J. Sci. Eng. Res*, vol. 3, pp. 1–4, 2012.

[5] G. Dillingham, "Aviation security: Weaknesses in airport security and options for assigning screening responsibilities," US General Accounting Office, Tech. Rep., 2001.

[6] B. Elias, "Airport passenger screening: Background and issues for congress," Tech. Rep., 2010.

[7] S. Naik and C. Murthy, "Hue-preserving color image enhancement without gamut problem," *IEEE Transactions on image processing*, vol. 12, no. 12, pp. 1591–1598, 2003.

[8] O. Dorgham, I. Al-Mherat, J. Al-Shaer, S. Bani-Ahmad, and S. Laycock, "Smart system for prediction of accurate surface electromyography signals using an artificial neural network," *Future Internet*, vol. 11, no. 1, 2019.

[9] K. Kase, "Effective use of color in x-ray image enhancement for luggage inspection," Master's thesis, Master's Degree thesis, 2002.

[10] C. Blehm, S. Vishnu, A. Khattak, S. Mitra, and R. Yee, "Computer vision syndrome: a review," *Surv Ophthalmol*, vol. 50, no. 3, pp. 253–262, 2005.

[11] Y. Sadah, N. Al-Najdawi, and S. Tedmori, "Exploiting hybrid methods for enhancing digital x-ray images," *Int. Arab J. Inf. Technol.*, vol. 10, no. 1, pp. 28–35, 2013.

[12] M. Fisher, O. Dorgham, and S. Laycock, "Fast reconstructed radiographs from octree-compressed volumetric data," *Int J Comput Assist Radiol Surg*, vol. 8, no. 2, 2013.

[13] P. Głomb, M. Romaszewski, M. Cholewa, and K. Domino, "Application of hyperspectral imaging and machine learning methods for the detection of gunshot residue patterns," *Forensic Sci Int*, vol. 290, pp. 227–237, 2018.

[14] Z. Liu, J. Li, Y. Shu, and D. Zhang, "Detection and recognition of security detection object based on yolo9000," in *2018 5th International Conference on Systems and Informatics (ICSAI)*. IEEE, 2018, pp. 278–282.

[15] L. Anthopoulos and V. Kazantzi, "Urban energy efficiency assessment models from an ai and big data perspective: tools for policy makers," *Sustain Cities Soc*, vol. 76, p. 103492, 2022.

[16] S. Mehta, B. Bhushan, and R. Kumar, "Machine learning approaches for smart city applications: Emergence, challenges and opportunities," in *Recent Advances in Internet of Things and Machine Learning*, 2022, pp. 147–163.

[17] G. Chen and J. Mehta, "Applying artificial intelligence and deep belief network to predict traffic congestion evacuation performance in smart cities," *Appl Soft Comput*, vol. 121, p. 108692, 2022.

[18] J. Guo and Z. Lv, "Application of digital twins in multiple fields," *Multimed Tools Appl*, pp. 1–27, 2022.

[19] S. Chackravarthy, S. Schmitt, and L. Yang, "Intelligent crime anomaly detection in smart cities using deep learning," in *2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC)*. IEEE, 2018, pp. 399–404.

[20] A. Baughman, C. Eggenberger, A. Martin, D. Stoessel, and C. Trim, "Incident prediction and response using deep learning techniques and multimodal data," Patent, mar, 2022, uS Patent App. 16/832,021.

[21] S. Jain, M. Alam, G. Kurubacak, and N. Labib, "Opportunities and challenges of cyber-physical transportation systems," in *Handbook of Research of Internet of Things and Cyber-Physical Systems*, 2022, pp. 579–602.

[22] N. Hussein, F. Hu, and F. He, "Multisensor of thermal and visual images to detect concealed weapon using harmony search image fusion approach," *Pattern Recognit Lett*, vol. 94, pp. 219–227, 2017.

[23] R. Olmos, S. Tabik, A. Lamas, F. Pérez-Hernández, and F. Herrera, "A binocular image fusion approach for minimizing false positives in handgun detection with deep learning," *Information Fusion*, vol. 49, pp. 271–280, 2019.

[24] S. Dubey, "Building a gun detection model using deep learning," in *Program Chair & Proceedings Editor: M. Afzal Upal, PhD Chair of Computing & Information Science Department Mercyhurst University*, vol. 501, 2019.

[25] A. Matiolański, A. Maksimova, and A. Dziech, "Cctv object detection with fuzzy classification and image enhancement," *Multimed Tools Appl*, vol. 75, no. 17, pp. 10 513–10 528, 2016.

[26] S. Akcay, M. Kundegorski, C. Willcocks, and T. Breckon, "Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery," *IEEE transactions on information forensics and security*, vol. 13, no. 9, pp. 2203–2215, 2018.

[27] M. Kowalski, M. Kastek, M. Piszczek, M. Życzkowski, and M. Szustakowski, "Harmless screening of humans for the detection of concealed objects," in *Safety and Security Engineering VI*, vol. 151, 2015, pp. 215–223.

[28] N. Nandhini, M. Sowmya, and P. Malarvizhi, "Spontaneous detection of weapons in atm banks," *IJEDR*, vol. 5, no. 1, 2017.

[29] A. Goyal *et al.*, "Automatic border surveillance using machine learning in remote video surveillance systems," in *Emerging Trends in Electrical, Communications, and Information Technologies*. Springer, 2020, pp. 751–760.

[30] 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.

[31] O. Dorgham, M. Naser, M. Ryalat, A. Hyari, N. Al-Najdawi, and S. Mirjalili, "U-netcts: U-net deep neural network for fully automatic segmentation of 3d ct dicom volume," *Smart Health*, vol. 26, p. 100304, 2022.

[32] O. Dorgham, "Automatic body segmentation from computed tomography image," in *Proceedings - 3rd International Conference on Advanced Technologies for Signal and Image Processing, ATSIP 2017*, 2017.

[33] F. Gelana and A. Yadav, "Firearm detection from surveillance cameras using image processing and machine learning techniques," in *Smart innovations in communication and computational sciences*, 2019, pp. 25–34.

[34] J. Lai and S. Maples, "Developing a real-time gun detection classifier, world academy of science," *Stanford University*, 2017.

[35] M. K. E. Den, A. Taha, and H. H. Zayed, "Automatic gun detection approach for video surveillance," *International Journal of Sociotechnology and Knowledge Development (IJSKD)*, vol. 12, no. 1, pp. 49–66, 2020.

[36] S. Aburass and O. Dorgham, "Performance evaluation of swin vision transformer model using gradient accumulation optimization technique," in *Proceedings of the Future Technologies Conference*, 2023, pp. 56–64.

[37] Z. Xue and R. Blum, "Concealed weapon detection using color image fusion," in *Proceedings of the 6th international conference on information fusion*, vol. 1. IEEE, 2003, pp. 622–627.

[38] A. Achanta, M. McKenna, and J. Heyman, "Nonlinear acoustic concealed weapons detection," in *34th Applied Imagery and Pattern Recognition Workshop (AIPR'05)*. IEEE, 2005, pp. 7–11.

[39] M. Elsdon, D. Smith, M. Leach, and S. Foti, "Microwave imaging of concealed metal objects using a novel indirect holographic method," *Microw Opt Technol Lett*, vol. 47, no. 6, pp. 536–537, 2005.

[40] S. Kolkoori, N. Wrobel, K. Osterloh, U. Zscherpel, and U. Ewert, "Novel x-ray backscatter technique for detection of dangerous materials: application to aviation and port security," *Journal of Instrumentation*, vol. 8, no. 09, p. P09017, sep 2013. [Online]. Available: https://dx.doi.org/10.1088/1748-0221/8/09/P09017

[41] O. Dorgham, "Performance of a 2d-3d image registration system using (lossy) compressed x-ray ct," *Annals of the BMVA Vol*, vol. 2008, no. 3, pp. 1–11, 2008.

[42] O. Dorgham, M. Ryalat, and M. Naser, "Automatic body segmentation for accelerated rendering of digitally reconstructed radiograph images," *Inform Med Unlocked*, vol. 20, p. 100375, 2020.

[43] R. Gonzalez, *Digital image processing*. Pearson Education India, 2009.

[44] N. Al-Najdawi, S. Tedmori, O. Alzubi, O. Dorgham, and J. Alzubi, "A frequency based hierarchical fast search block matching algorithm for fast video communication," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 4, 2016.

[45] Y. Freund and R. Schapire, "Experiments with a new boosting algorithm," in *icml*, vol. 96. Citeseer, 1996, pp. 148–156.

[46] Y. F and R. 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.

[47] W.-C. Cheng and D.-M. Jhan, "Triaxial accelerometer-based fall detection method using a self-constructing cascade-adaboost-svm classifier," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 2, pp. 411–419, 2012.

[48] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, vol. 1. IEEE, 2001, pp. I–I.

[49] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[50] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[51] A. Sophian, G. Tian, D. Taylor, and J. Rudlin, "A feature extraction technique based on principal component analysis for pulsed eddy current ndt," *NDT & E International*, vol. 36, no. 1, pp. 37–41, 2003.

[52] M. Li and B. Yuan, "2d-lda: A statistical linear discriminant analysis for image matrix," *Pattern Recognition Letters*, vol. 26, no. 5, pp. 527–532, 2005.

[53] M. Bartlett, J. Movellan, and T. Sejnowski, "Face recognition by independent component analysis," *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1450–1464, 2002.

[54] D. Gabor, "Theory of communication. part 1: The analysis of information," *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[55] D. Nguyen, Z. Zong, P. Ogunbona, and W. Li, "Object detection using non-redundant local binary patterns," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 4609–4612, 2010.

[56] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[57] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[58] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1635–1650, 2010.

[59] N. Sun, W. Zheng, C. Sun, C. Zou, and L. Zhao, "Gender classification based on boosting local binary pattern," in *International Symposium on Neural Networks*. Springer, 2006, pp. 194–201.

[60] Z. Yang and H. Ai, "Demographic classification with local binary patterns," in *International Conference on Biometrics*. Springer, 2007, pp. 464–473.

[61] T. Gritti, C. Shan, V. Jeanne, and R. Braspenning, "Local features based facial expression recognition with face registration errors," in *2008 8th IEEE International Conference on Automatic Face & Gesture Recognition*. IEEE, 2008, pp. 1–8.

[62] B. Ma, W. Zhang, S. Shan, X. Chen, and W. Gao, "Robust head pose estimation using lgbp," in *18th International Conference on Pattern Recognition (ICPR'06)*. IEEE, 2006, pp. 512–515.

[63] X. Wei, H.-M. Chen, and I. Amad, "Millimeter-wave video sequence denoising and enhancement in concealed weapons detection application," in *Applications of Digital Image Processing XXIX*, vol. 6312. SPIE, 2006, pp. 55–62.

[64] D. Saavedra, S. Banerjee, and D. Mery, "Detection of threat objects in baggage inspection with x-ray images using deep learning," *Neural Computing and Applications*, vol. 33, pp. 7803–7819, 2021. [Online]. Available: https://doi.org/10.1007/s00521-020-05521-2

Evaluating Transparency in the Development of Artificial Intelligence Systems: A Systematic Literature Review

Giulia Karanxha, Paulinus Ofem
Business and Information Technology, Laurea University of Applied Sciences, Espoo, Finland

***Abstract*—Transparency is increasingly recognised as a cornerstone of trustworthy artificial intelligence (AI), yet its operationalisation remains fragmented and underdeveloped. Existing methods often rely on qualitative checklists or domain-specific case studies, limiting comparability, reproducibility, and regulatory alignment. This paper presents a Systematic Literature Review (SLR) of 28 peer-reviewed studies that explicitly propose or apply methods for evaluating transparency in AI systems (2019-July 2025). The review identifies recurring themes such as traceability, explainability, and communication, and classifies evaluation approaches by metric type and calculation type. Empirically, checklist-based instruments are the most frequent evaluation form (9/28, 32%), followed by scenario-based qualitative assessments (5/28, 18%). Most (9/28, 32%) research on AI applications occurs in healthcare; references to legal or ethical frameworks appear in 19/28 studies (67%), although traceable mappings to specific obligations are rare. The results of the quality assessment highlight strengths in methodological clarity, but reveal persistent gaps in benchmarking, stakeholder inclusion, and lifecycle integration. Based on these findings, this study informs the adaptation of the Z-Inspection® process within the context of AI development projects and motivates a Transparency Artefact Registry (TAR), a structured, metadata-based mechanism for capturing and reusing transparency artefacts across system lifecycles. By embedding transparency evaluation into AI development workflows, the proposed approach seeks to provide verifiable, repeatable, and regulation-aligned practices for assessing transparency in complex AI systems.**

Keywords*—Artificial intelligence; transparency evaluation; trustworthy AI; transparency metrics; EU AI Act; systematic literature review

I. Introduction

The rapid deployment of artificial intelligence (AI) in high-impact domains has intensified calls for transparency as a cornerstone of Trustworthy AI. Transparency enables stakeholders to understand, contextualise, and, where necessary, contest AI-driven decisions [1]. Its importance is underscored by the EU Artificial Intelligence Act (Regulation (EU) 2024/1689), which mandates that high-risk AI systems and general-purpose AI models provide features such as explainability, traceability, and clear user information [2]. Yet, despite these obligations, the operationalisation of transparency remains fragmented, with varying definitions, scopes, and evaluation practices across disciplines and application domains. This poor clarity risks undermining both accountability and compliance in real-world AI deployments.

For the MANOLO Project (Trustworthy Efficient AI for Cloud-Edge Computing) [3], this challenge presents an opportunity to explore further. This project aims to develop a complete set of trustworthy algorithms and tools that would enable AI systems to achieve better efficiency and optimise their operations on the edge of the cloud continuum and in sectors such as healthcare, telecommunications, and robotics, where systems must combine performance and energy requirements with strict ethical and legal compliance. In such contexts, the absence of structured, measurable approaches to transparency achievement and evaluation poses not only a compliance risk but also a practical barrier to demonstrating accountability to regulators, end-users, and other stakeholders.

To address this, we conducted a Systematic Literature Review (SLR) of peer-reviewed studies published between January 2019 and July 2025. The review identifies how transparency in AI is defined, measured, and operationalised, and where gaps remain. By mapping existing approaches and their limitations, the SLR provides the empirical and conceptual foundation needed for AI projects to adopt a structured transparency evaluation strategy.

This review also serves as the basis for further methodological contributions. Its findings directly inform our conference paper, which introduced an artefact-driven approach to transparency evaluation in applied AI contexts. In this vein, this paper is not an isolated study but the first step in a broader research programme: establishing a systematic evidence base for transparency evaluation, and subsequently proposing solutions that can be embedded into frameworks such as Z-Inspection® [4] to support continuous, auditable, and stakeholder-relevant transparency in AI projects and beyond.

The remainder of this paper is organised as follows. Motivation situates the work in the MANOLO and Z-Inspection® contexts; Related Work maps policy, governance, and SE/IS perspectives; Methodology details the SLR protocol (scope, PICOC, search, screening, quality appraisal, extraction); Results report the evidence base and descriptive statistics; Thematic Analysis synthesises cross-cutting patterns; Discussion provides implications of findings and advantages over prevalent approaches; Conclusion summarises limitations and future work.

Contributions and roadmap. This study 1) focuses on operational, metric-based transparency *evaluation* methods rather than principles, 2) proposes a compact comparative taxonomy (metric types) and quantifies patterns across 28 studies, 3) offers a cross-domain synthesis bridging AI governance and SE/IS consideration of transparency as a first class non-functional requirement, and 4) links identified gaps to an

artefact-driven solution space (Transparency Artefact Registry, TAR) aligned with Z-Inspection® and the EU Artificial Intelligence Act (EU AI Act). Section II motivates the challenge of evaluating transparency in AI development projects such as the Manolo project; Section III synthesises related work; Section IV details the SLR protocol; Section V reports descriptive results; Section IV-G provides thematic analysis; Section VII discusses implications; Section VIII concludes.

II. Motivation

Within MANOLO, the evaluation of AI trustworthiness is grounded in the Z-Inspection® framework [4], a holistic process for assessing AI's ethical compliance, technical robustness, and legal alignment. Although Z-Inspection® provides comprehensive coverage of the trustworthiness dimensions, it does not offer a structured, dedicated methodology for the evaluation of transparency throughout the AI lifecycle.

For AI projects that utilise Z-Inspection®, this limitation creates both a compliance risk under the EU AI Act and a practical challenge to demonstrate transparency and accountability to stakeholders. Therefore, a systematic investigation of existing transparency evaluation methods was required to establish a structured evidence base. This SLR responds to this need by synthesising how transparency has been defined, measured, and operationalised in peer-reviewed studies. In this way, the review provides the foundation for future AI projects that will adapt Z-Inspection® with measurable transparency criteria aligned with regulatory obligations and stakeholder needs.

III. Related Work

Transparency is increasingly recognised as a key requirement for ethical and trustworthy AI, yet its practical achievement and evaluation remain challenging. Over 94% of ethical AI guidelines include transparency or explainability as core principles [5], but few provide actionable tools for assessment. This gap between normative commitments and operational practices has driven attention from regulators, practitioners, and researchers alike.

A. Policy and Regulatory Frameworks

The concept of transparency gained formal traction in the European policy landscape through the High-Level Expert Group on AI (AI HLEG), which identified transparency, encompassing traceability, explainability, and user communication, as one of seven requirements for Trustworthy AI [1]. While influential in shaping the EU AI Act [2], these guidelines are not legally binding. The Act introduces layered obligations for high-risk AI systems (Article 13), general-purpose AI models, and AI systems that interact with humans (Article 50), but leaves their operationalisation open to interpretation [6].

Sector-specific initiatives further illustrate this shift from high-level principles to more practical instruments. The TITAN Guideline [7] offers a checklist for AI disclosure in scholarly publishing, while Genovesi et al. [8] compare how transparency is addressed in standards and policy artefacts. Although valuable, such approaches remain largely domain-bound and rarely demonstrate applicability across diverse AI systems.

B. Societal Accountability and Governance

Beyond regulatory discourse, concerns persist about the societal implications of opaque AI. Busuioc [9] underscores the risks of algorithmic systems becoming embedded in the public sector while remaining resistant to challenge, thereby undermining institutional accountability. This perspective positions transparency not only as a compliance issue but also as a prerequisite for democratic oversight. In line with this, Panigutti et al. [10] argue that transparency must be approached as a dynamic, stakeholder-aware practice rather than a static disclosure requirement. These insights highlight the broader governance role of transparency, which extends beyond technical explainability.

C. Software Engineering and Information Systems Foundations

Long before the rise of AI governance, transparency was conceptualised within software engineering (SE) and information systems (IS) as a non-functional requirement (NFR). Leite and Cappelli [11] introduced transparency as a "soft goal," supported by attributes such as accessibility, auditability, and usability. Hosseini et al. [12] extended this by framing transparency as a bidirectional process between information providers and receivers, thereby emphasising the importance of communication and user understanding. Spagnuelo et al. [13] operationalised the concept further by proposing sector-specific metrics for auditability, accountability, and verifiability in IT systems.

Building on these conceptual foundations, Ofem and colleagues advanced transparency in software engineering through a sequence of works. Their 2022 systematic review [14] clarified the fragmented definitions of transparency and positioned it explicitly as a non-functional requirement, comparable to established quality attributes such as reliability or usability. In the same year, Isong, Ofem, and Lugayizi [15] proposed a framework that operationalises transparency via quality factors, such as accessibility, usability, understandability, modifiability, and reusability, each linked to artefact- and process-level indicators. While these factors provide a structured starting point, the framework acknowledges that specific instantiations may be domain-dependent and must be adapted to stakeholder requirements. This line of work culminated in the 2024 metrics study [16], which introduced empirically validated measurement instruments and an improvement model, thereby transforming transparency from an abstract principle into a set of actionable, lifecycle-aware practices. Although developed in the SE domain, these contributions are directly relevant to AI governance, where similar challenges of definition, measurement, and lifecycle integration remain unresolved.

D. Relevance to AI Transparency Evaluation

Although originating in SE and IS, these approaches are directly relevant to AI governance. The challenges of lifecycle integration, stakeholder involvement, and continuous evaluation are not unique to AI but are shared with complex software systems more broadly. Yet, AI-focused frameworks have thus far incorporated these traditions only partially. Most emphasise explainability or documentation, but rarely adopt SE/IS practices such as artefact registries, maturity models, or quality factor-metric mappings.

This review therefore situates AI transparency evaluation within a broader trajectory: from ethical and regulatory guidelines, through societal accountability concerns, to structured SE/IS methodologies. While progress has been made in translating abstract principles into operational criteria, most existing efforts remain either domain-specific, non-standardised, or lacking in mechanisms for continuous lifecycle monitoring. The present SLR contributes to addressing that gap by systematically identifying and analysing peer-reviewed studies that explicitly propose or apply methods for evaluating transparency in AI systems, and by assessing their relationship to established SE/IS traditions.

1) Integrative synthesis: Policy and governance sources prioritise obligations, rights, and auditability, which have produced tools that mostly take the form of checklists and high-level controls. In contrast, SE/IS literature frames transparency as a non-functional requirement and develops artefact and process-level indicators with measurable quality factors. These viewpoints converge around traceability and documentation, taking into account logs, records, and provenance as core evidence, while also acknowledging lifecycle concerns. These perspectives diverge in emphasis. AI governance often treats explainability as a proxy for transparency, while SE/IS considers transparency as a concept with measurable quality constructs. Methodological overlaps lack coherence: many governance-oriented frameworks document evidence without scoring, whereas several SE/IS approaches score internal artefacts yet rarely assess whether information is communicated to stakeholders in clear, timely, and usable ways (e.g., plain-language notices, model cards, user-tested interfaces). Taken together, this reveals the gap this SLR targets: the field lacks a standardised, lifecycle-grounded, cross-domain evaluation scheme that connects regulatory obligations to measurable artefacts and metrics.

IV. Methodology

This study adopts a SLR approach to investigate how transparency is evaluated in artificial intelligence (AI) systems. The review is guided by the following research questions:

- RQ1: What are the existing approaches for implementing and evaluating transparency in AI systems?
- RQ2: How do these approaches address key transparency concerns?
- RQ3: What are the strengths, limitations, and gaps of these approaches?
- RQ4: How applicable are these approaches in real-world AI systems?

The aim is to identify methodological trends, assess sectoral and regulatory relevance, and support the development of actionable stakeholder-centred transparency evaluation strategies.

TABLE I. Total Search Results Summary

Engine	Total	Reviewed	Overlap
Google Scholar	45,700	179	9
IEEE Xplore	1,799	83	–
Web of Science	4,341	104	22
ScienceDirect	3,998	105	2
ACM DL	4,078	71	7

A. Scope Framing and Objectives

Transparency was examined as a multidimensional concept that encompasses traceability, communication, and explainability [17]. The purpose of the review was to synthesise existing evaluation methods, identify common practices and limitations, and provide a structured basis for future development of transparency metrics and frameworks. The review focused on peer-reviewed literature published between January 2019 and July 2025 to reflect the growing influence of transparency in AI governance discourse, especially in light of regulatory developments such as the EU AI Act (2024) [2].

B. PICOC Structure

The scope of the review was defined using the PICOC model.

- Population: AI systems deployed across various sectors and domains.
- Intervention: Methods, frameworks, or metrics introduced or applied to evaluate transparency.
- Comparison: different approaches for implementing and evaluating transparency.
- Outcome: Assessment of the strengths, limitations, and applicability of approaches to AI development projects.
- Context: AI development projects.

This framing supported a focused yet cross-sectoral exploration of how transparency is operationalised and assessed in contemporary AI research.

C. Search Strategy

The literature search was conducted between January 2019 and July 2025 across five major academic databases: Google Scholar, IEEE Xplore, Web of Science, ScienceDirect, and the ACM Digital Library. These sources were selected for their broad coverage of AI, computer science, ethics, and governance literature. The consolidated counts are summarised in Table I. Search queries combined keywords related to transparency, artificial intelligence, and evaluation, adapted to the syntax of each platform. An example query was:

```
("Artificial intelligence system" OR "AI
        software") AND
```

```
("AI transparency" OR "transparent AI" OR
"trustworthy AI") AND
("transparency evaluation" OR
"transparency metrics" OR "transparency
framework" OR "trustworthy framework" OR
"assessment")
```

Boolean operators and filters (e.g., English language, peer-reviewed) were applied where supported. A heuristic stopping rule was used to manage feasibility, with searches concluded when ten consecutive irrelevant results were encountered. A total of over 59,000 results were screened, including 45,700 from Google Scholar, 4,341 from Web of Science, 1,799 from IEEE, 4,078 from ACM, and 3,998 from ScienceDirect.

D. Screening and Eligibility (PRISMA)

Screening followed the PRISMA 2020 framework. The selection and exclusion process is summarised in the PRISMA flow diagram shown in Fig. 1. After duplicate removal and title/abstract screening, 542 articles were assessed. A total of 28 peer-reviewed studies were included in the final review. The inclusion and exclusion criteria were:

- Inclusion:
  - Peer-reviewed journal or conference publications
  - Published between January 2019 and July 2025
  - Explicitly introduced or applied a method, framework, or metric to evaluate transparency in AI systems
- Exclusion:
  - Discussed transparency only conceptually, without operational methods or proposed governance mechanisms
  - Focused solely on explainability, fairness, or accountability without reference to transparency
  - Non-English or inaccessible full texts

E. Quality Assessment

The included studies were evaluated using a five-criterion quality checklist, adapted from a prior systematic literature review in software engineering [14]:

- Is the evaluation method explicitly described?
- Are structured criteria or metrics provided?
- Are the strengths and limitations discussed?
- Is the method compared with other frameworks?
- Is the approach applied in real-world AI use cases?

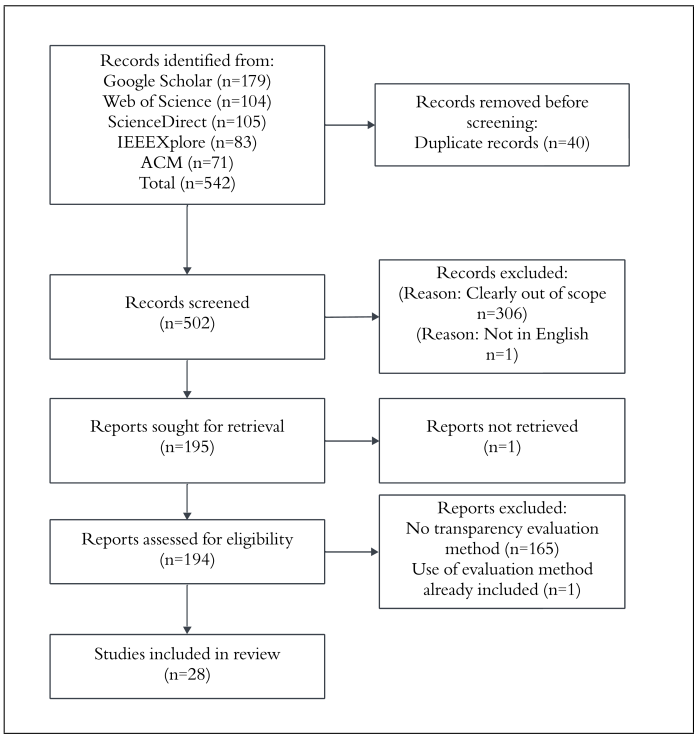


Fig. 1. PRISMA 2020 screening flow for the review window January 2019-July 2025.

Each item was scored as Yes (1), Partly (0.5), or No (0). Studies were not excluded based on their quality score, but scores were used descriptively to support analysis. As shown in Table II, most studies scored 4 or above out of 5.

F. Data Extraction and Classification

A structured extraction template was applied to each study, capturing:

- Transparency definition and theoretical framing.
- Name and structure of evaluation framework or method.
- Metric type (e.g., checklist, score, index, rating, qualitative, survey, explainability-based).
- Metric calculation type (e.g., single score, ordinal classification, aggregated, scenario-based).
- Sector or domain of application.
- Alignment with legal or ethical frameworks (e.g., GDPR, EU AI Act).
- Reported strengths and limitations.

Metrics were classified using a two-dimensional taxonomy: metric type and calculation type. Where multiple techniques

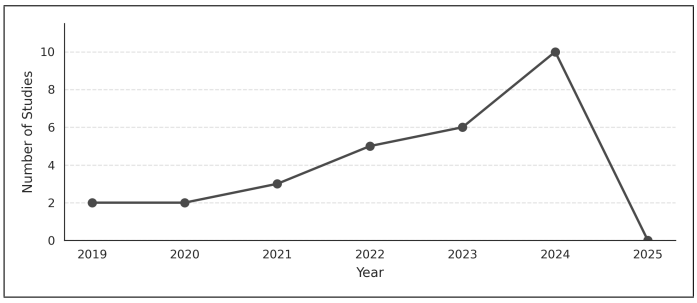


Fig. 2. Included studies per year (2019-2025; cut-off July 2025), n=28.

were reported, the dominant method was selected based on its evaluative function.

1) Data management: Extracted variables were recorded in a Microsoft Excel workbook, which constitutes the coded analytic dataset used for the descriptive analyses and quality appraisal.

2) Dataset: No experimental or task datasets were analysed. The only data used in this study are the 28 included papers (2019-July 2025) and the coded Excel extraction sheet derived from them. The Excel sheet will be made available upon request.

G. Thematic Analysis

Thematic synthesis followed Braun and Clarke's method [18]. Codes were generated from extracted data fields and grouped inductively into higher-level themes. These themes reflect patterns in how transparency is conceptualised and evaluated across the selected studies. The analysis served to uncover recurrent gaps and identify emerging evaluation practices.

H. Limitations

Several limitations should be acknowledged. First, the review covered the period from January 2019 to July 2025 and focused on peer-reviewed English-language literature, excluding grey literature and non-English sources. Second, the quality assessment involved a degree of researcher judgment, particularly in borderline cases. Third, while metric types and calculation methods were standardised to support comparison, some hybrid approaches may not have been fully captured by this classification. Lastly, the inclusion criteria required that transparency be explicitly evaluated, which may have led to the exclusion of relevant studies where transparency was subsumed under broader constructs such as fairness or trustworthiness.

V. Results

This section presents the results of the systematic literature review, providing a descriptive overview of the selected studies, including the classification of evaluation methods, the distribution of metric types and calculation approaches, as well as the sectoral focus, implementation status, and references to relevant legal and ethical frameworks.

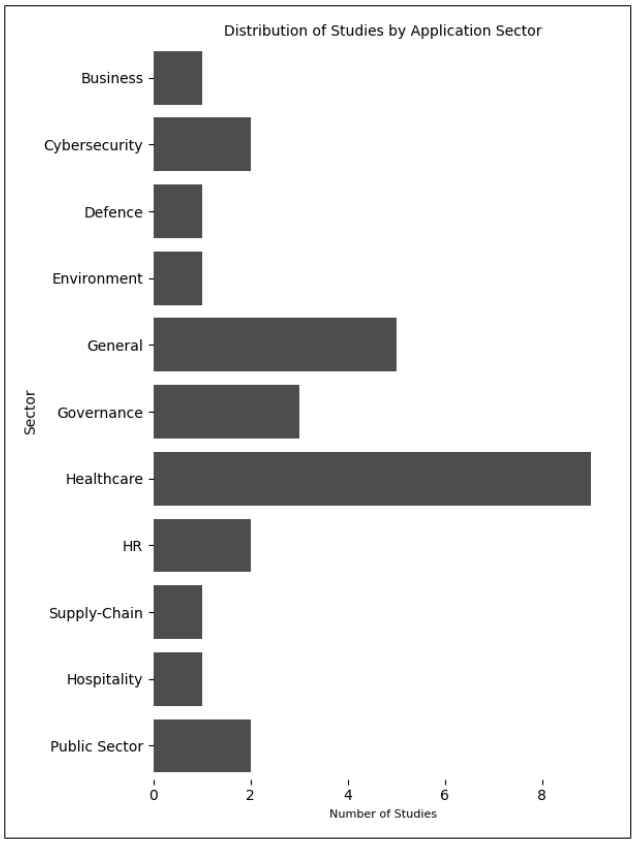


Fig. 3. Sector distribution of the 28 included studies; healthcare 9/28 (32%).

A. Overview of Included Studies

A total of 28 peer-reviewed studies published between January 2019 and July 2025 were selected for inclusion. The majority introduced formal frameworks or tools aimed at evaluating transparency in artificial intelligence systems, while a smaller subset proposed conceptual approaches grounded in governance or policy-oriented perspectives. In terms of application status, most of the studies (23 out of 28) applied their method to a real-world system, dataset, or use case. At the same time, the rest remained conceptual, theoretical, or exploratory in scope. See Fig. 2 for the year-by-year distribution.

The analysis revealed a notable concentration of transparency evaluation in the healthcare sector, with 9 of the 28 studies (32%) focusing on clinical decision support, medical imaging, or digital health systems [19], [20], [21], [4], [22], [23], [24], [25], [26]. Other studies proposed general-purpose frameworks or were situated in domains such as cybersecurity, HR, or public sector applications. A smaller number of studies addressed domains such as hospitality or defence. This sectoral imbalance highlights the strong regulatory and ethical pressures in healthcare, in contrast to more limited methodological development in other fields. See Fig. 3 for sectoral concentration.

B. Metric Types

The studies were categorised according to the type of metric or evaluation logic employed. Seven primary metric

types were identified:

- Checklist-based approaches operationalise transparency through binary questions or structured assessment lists. [27][19][28][25][29][30][17][31][20]
- Qualitative methods rely on expert judgement or narrative analysis. [21][4][32][33][34]
- Explainability-based evaluations assess transparency through interpretability outputs (e.g., SHAP, LIME). [35][36][37][38]
- Survey-based tools capture perceptions of transparency using structured questionnaires. [39][22][40][41]
- Index-based approaches produce composite scores or traceability indexes. [26][24][42]
- Score-based methods generate transparency scores by assigning weighted values to defined criteria. [43][23]
- Rating-based frameworks. [44]

Checklist-based approaches were the most common, often reflecting broader trends in AI governance tools such as the ALTAI framework [17].

C. Regulatory Alignment

References to legal and ethical frameworks were present in 19 of the 28 studies (67%) [19], [17], [31], [20], [4], [39], [22], [44], [33], [27], [34], [36], [42], [28], [24], [25], [38], [26], [29], most commonly the EU Artificial Intelligence Act, the General Data Protection Regulation (GDPR), and the AI HLEG Guidelines. However, only a few studies provided traceable mappings between evaluation criteria and specific legal provisions. For instance, RETENTION [26] and POLARIS [28] embedded elements of the EU AI Act [2] and AI HLEG recommendations [17] within their structure. In the majority of cases, alignment remained high-level or aspirational, with limited support for formal compliance, certification, or auditability.

D. Quality Assessment

Across the 28 studies, the overall quality was high, with most achieving at least 70% of the maximum possible score (see Table II). Several studies attained full or nearly full scores, while others performed lower due to conceptual or preliminary scope.

Specifically, QCA1 (evaluation method outlined) scored the highest, achieving 96.4% of the maximum score. This outcome was expected, as the presence of an explicitly described transparency evaluation method was a prerequisite for inclusion during the full-text review. QCA3 (discussion of strengths and limitations) also performed strongly, with 94.6%. By contrast, QCA2 (specific criteria for transparency assessment) and QCA5 (use of real-world AI applications) achieved 76.8% each, indicating room for improvement in applying structured criteria and real-world validation. The lowest-performing criterion was QCA4 (comparison with other transparency frameworks), with only 46.4% of the maximum score, showing that relatively few studies benchmarked their approach against other methods.

As shown in Table II, studies with higher total scores were typically more detailed, provided stronger methodological justification, and included real-world applications. In contrast, lower-scoring studies tended to be more conceptual or exploratory. These quality assessment results guided the interpretation of the final analysis, following best practices from prior systematic reviews (e.g., [14]), and support a nuanced discussion of strengths and limitations reported in Section Thematic Analysis.

VI. Thematic Analysis

This section presents the thematic analysis of the 28 studies included in the review. The analysis identified six overarching themes that characterise how transparency is defined, operationalised, and evaluated in artificial intelligence systems. These themes reflect both recurring strengths and persistent limitations across the literature.

A. Theme 1: Transparency as a Multi-Dimensional Construct

A consistent finding across the reviewed studies is the conceptualisation of transparency as a composite of multiple, interrelated dimensions. This mirrors the influential framing of the High-Level Expert Group on AI, which identified transparency, encompassing traceability, explainability, and communication, as one of the seven key requirements for Trustworthy AI [1]. These three pillars provided an early normative anchor and were frequently echoed, either explicitly or implicitly, in the SLR dataset.

Traceability concerns the documentation, logging, and oversight of system development, data flows, and decision-making processes, enabling stakeholders to follow and verify AI behaviour across its lifecycle. Explainability focuses on the interpretability of models or outputs, employing techniques such as SHAP, LIME, or saliency maps to make system reasoning intelligible. Communication refers to the disclosure of relevant information to users, regulators, or affected individuals in ways that are accessible, usable, and contextually appropriate.

While most studies referenced all three pillars, few operationalised them in an integrated manner. For example, IntelliLung [24], RETENTION [26], and SMACTR [30] explicitly combined traceability mechanisms, explainability techniques, and user-facing communication strategies within a single evaluation framework. In contrast, most approaches prioritised one or two dimensions, with communication often underdeveloped or absent. This imbalance reflects a gap in translating high-level principles into holistic evaluation designs.

This multi-dimensional perspective aligns closely with the stakeholder-centred framework of Ofem et al. [16], where

TABLE II. QUALITY ASSESSMENT RESULTS FOR INCLUDED STUDIES (N=28)

Study ID	QCA1	QCA2	QCA3	QCA4	QCA5	Total	% of Max
PS1	1	1	1	0.5	1	4.5	90%
PS2	1	0.5	0.5	0.5	0	2.5	50%
PS3	1	0.5	1	0.5	1	4.0	80%
PS4	1	0.5	1	0	1	3.5	70%
PS5	1	1	0.5	0	0	2.5	50%
PS6	1	0.5	1	0	0.5	3.0	60%
PS7	1	1	1	1	0	4.0	80%
PS8	1	0.5	1	0.5	1	4.0	80%
PS9	1	0.5	1	0.5	1	4.0	80%
PS10	1	0.5	1	0	1	3.5	70%
PS11	1	1	1	0.5	1	4.5	90%
PS12	1	1	1	1	1	5.0	100%
PS13	0.5	0.5	1	1	0	3.0	60%
PS14	1	1	0.5	0.5	1	4.0	80%
PS15	1	1	1	0.5	1	4.5	90%
PS16	1	1	1	0.5	1	4.5	90%
PS17	0.5	0.5	1	0	0.5	2.5	50%
PS18	1	1	1	0.5	1	4.5	90%
PS19	1	0.5	1	0.5	1	4.0	80%
PS20	1	1	1	0.5	1	4.5	90%
PS21	1	0.5	1	0.5	1	4.0	80%
PS22	1	1	1	0.5	0	3.5	70%
PS23	1	0.5	1	1	1	4.5	90%
PS24	1	1	1	0.5	1	4.5	90%
PS25	1	0.5	1	0	1	3.5	70%
PS26	1	1	1	0	0.5	3.5	70%
PS27	1	1	1	1	1	5.0	100%
PS28	1	1	1	0.5	1	4.5	90%
Total	27	21.5	26.5	13	21.5	109.5	—
% of Max	96.4%	76.8%	94.6%	46.4%	76.8%	—	—

transparency is linked to quality factors such as accessibility, usability, and understandability. Mapping the three pillars to these factors revealed that traceability overlaps most strongly with accessibility and modifiability, explainability with understandability, and communication with accessibility and usability. This mapping provides a bridge between AI ethics discourse and measurable software engineering practices.

In the context of this SLR, the prevalence of these three pillars directly addresses RQ1 (existing approaches to transparency evaluation) by highlighting their role as foundational building blocks. However, their uneven treatment across studies also informs RQ3 (strengths, limitations, and gaps), underscoring the need for integrated approaches that treat traceability, explainability, and communication as mutually reinforcing rather than isolated components.

B. Theme 2: Dominance of Checklists and Qualitative Methods

Across the 28 included studies, checklist-based approaches were the single most common way to evaluate transparency (*9/28; 32%*) [27], [19], [28], [25], [29], [30], [17], [31], [20], followed by scenario-based qualitative assessments (*5/28; 18%*) [21], [4], [32], [33], [34]. Explainability-output-driven evaluations (e.g., using SHAP/LIME artefacts) and survey instruments each appeared in *4/28* studies (*14%* apiece), while index-style composites were less frequent (*3/28; 11%*). Score-style metrics were rare (*2/28; 7%*), and explicit rating-scale schemes appeared only once (*1/28; 4%*).[1]

This pattern is consistent with the field's reliance on governance-aligned checklists (e.g., ALTAI-inspired instruments) and structured internal-audit playbooks (e.g., SMACTR; POLARIS; MLOps audit catalogues), which are accessible and broadly usable across domains [17], [30], [28], [25]. However, checklist implementations showed variable methodological quality in our review: while many were robust, some scored lower on our quality assessment (e.g., overall study QA scores in the 2.5-4.5 range), and, critically, few offered granular scoring, maturity levels, or CI/CD-friendly automation. In short, they structure the conversation, but they rarely quantify transparency in a way that is reproducible or benchmarkable across settings.

Significant counterexamples highlight how quantification can evolve. The Visibility Index Score (VIS) adapts supply-chain "visibility" logic to AI pipelines, scoring documentation quantity, freshness, and accuracy [43]. The Cyrus framework automates parts of transparency evaluation via FAIR-aligned metadata and linked-data quality tooling [23]. Both illustrate promising routes toward measurable, repeatable assessments—yet each remains bounded by domain and artefact conventions (VIS by reviewer judgment; Cyrus by metadata-rich, RDF/FAIR contexts). These limits inform our subsequent design choices: transitioning from static checklists to artefact-level metadata, evaluation linkages, and scorecard-driven gap closure that remain lightweight enough for real-world projects.

[1]These tallies refer to the dominant metric type per study when hybrids were present.

C. Theme 3: Limited Stakeholder and User Involvement

A striking gap in the reviewed literature is the limited inclusion of stakeholders, especially end-users, in the design, execution, and validation of transparency evaluations. Only a minority of studies (*6/28; 21%*) explicitly incorporated stakeholder perspectives into their methods, typically via structured interviews, surveys, or participatory workshops. Examples include IntelliLung [24], which involved clinicians to identify the most relevant artefacts and interpretability measures in a clinical decision-support context, and the ALTAI-inspired sectoral adaptations [27], [25], which consulted domain experts to weight checklist items. However, these engagements were generally one-off and did not establish feedback loops to iteratively refine the evaluation over the AI system lifecycle.

Despite regulatory emphasis on transparency being understandable to affected persons [2], only a small number of studies incorporate user perspectives into their evaluation frameworks. FAT-CAT [40] and the survey-based method by Fehr et al. [22] are exceptions, explicitly including perceptions of transparency, trust, and understanding in their measurement schemes.

A significant outlier to the lack of continuous stakeholder engagement is the Z-Inspection® Framework [4], which embeds a structured feedback loop across its Set-Up, Assess, and Resolve phases. While the stakeholders involved are often domain experts, project partners, or regulators rather than the ultimate end-users, the framework ensures that transparency requirements are revisited and refined after scenario analysis and ethical issue mapping, and that unresolved gaps feed into improvement actions. This iterative design demonstrates how multi-phase engagement can be operationalised, even if end-user representation remains limited.

From a software engineering perspective, the omission of such feedback mechanisms in most studies conflicts with stakeholder-driven models such as Ofem et al.'s transparency framework [16], which positions stakeholder requirements as the starting point for defining quality factors and associated metrics. Without ongoing stakeholder input, whether from end-users or intermediary stakeholders, transparency evaluations risk being misaligned with the information needs, literacy levels, and contextual constraints of those most affected by AI systems.

In relation to our research questions, this theme directly addresses RQ2 (how key transparency concerns are addressed) by showing that while stakeholder relevance is widely recognised in principle, it is rarely operationalised in a structured and continuous manner. It also informs RQ3 (strengths, limitations, and gaps), identifying the lack of sustained, multi-phase stakeholder participation as a significant limitation in current transparency evaluation practice.

D. Theme 4: Sectoral Concentration in Healthcare

Healthcare emerged as the dominant application domain for transparency evaluation frameworks in our SLR, with **9/28** studies (**32%**) focusing exclusively on medical AI systems. These ranged from diagnostic imaging and clinical decision-support tools to patient-facing health applications[19], [20], [21], [4], [22], [23], [24], [25], [26].

The concentration reflects healthcare's high regulatory scrutiny, established ethical oversight mechanisms, and the availability of structured artefacts such as clinical guidelines, audit logs, and model validation reports.

Relevant examples include IntelliLung [24], which applied an integrated traceability, explainability, and communication framework to assess an AI system for lung disease diagnosis, and RETENTION [26], which evaluated an AI tool for predicting patient readmission risks. Both demonstrated the feasibility of combining multiple transparency pillars with sector-specific artefacts and metrics, benefiting from healthcare's pre-existing documentation and audit cultures.

In contrast, sectors such as finance, transportation, and public administration were underrepresented, with most non-healthcare cases relying on adapted checklists (e.g., ALTAI) or scenario-based ethical assessments. These domains often lack the same level of artefact maturity or regulatory compulsion, resulting in evaluations that are either less comprehensive or highly context-dependent.

This imbalance has implications for RQ4 (applicability in real-world AI systems). While healthcare-focused frameworks offer valuable methodological depth, their artefact-rich environment may not translate directly to sectors where documentation practices are less formalised. For cross-sector applicability, transparency evaluation methods must adapt to different artefact availability, stakeholder profiles, and regulatory drivers without sacrificing rigour. This supports the need, identified in RQ3, for domain-flexible frameworks that can scale from artefact-rich to artefact-sparse environments.

E. Theme 5: Partial and Informal Regulatory Alignment

Many transparency evaluation approaches in the reviewed studies reference regulatory or ethical frameworks, most commonly the EU AI Act, the AI HLEG Ethics Guidelines for Trustworthy AI, or sector-specific standards. However, in *20/28* studies (*71%*), this alignment was *partial* or *informal*: regulatory documents were cited as conceptual anchors, but their specific obligations were not systematically translated into evaluation criteria or metrics.

For example, several healthcare-focused studies mapped their checklists to the AI HLEG requirements [17] or to medical device regulations, but did not provide a traceable mapping between regulatory clauses (e.g., EU AI Act Article 13 obligations for high-risk systems) and the artefacts or scores used in their evaluations. Similarly, frameworks such as ALTAI were adapted to specific organisational contexts without maintaining the whole structure or weighting defined in the original.

Only a small subset of studies (*4/28; 14%*) achieved strong formal alignment by creating explicit mappings from regulatory text to artefacts, evaluation metrics, and scoring logic. Examples include RETENTION [26] and POLARIS [28]. These approaches demonstrated clearer audit readiness and traceability of compliance claims but required significant up-front investment in legal-technical translation.

In relation to our research questions, this theme informs both RQ2 (how key transparency concerns are addressed) and RQ3 (strengths, limitations, and gaps). It shows that while

regulatory references are widespread, operationalisation into concrete, verifiable measures is inconsistent. This gap highlights the importance of evaluation frameworks that integrate regulatory requirements directly into artefact-metric mappings, allowing for compliance checks alongside broader ethical evaluations.

F. Theme 6: Gaps in Standardisation and Lifecycle Integration

A recurring limitation across the reviewed studies is the lack of standardisation in how transparency is defined, measured, and maintained over an AI system's lifecycle. Even when similar terminology is used (e.g., "traceability" or "model interpretability"), the underlying definitions, artefacts, and metrics vary considerably, making cross-study comparison and benchmarking difficult. This variability extends to the granularity of evaluation (system-level vs. component-level), the form of evidence accepted (qualitative descriptions vs. quantitative scores), and the frequency of assessment.

Lifecycle integration is also underdeveloped. Most frameworks focus on a single point-in-time evaluation, often predeployment or post-deployment, without mechanisms for reassessment as the system evolves. Continuous monitoring, when mentioned, was generally conceptual rather than operationalised through automated pipelines or version-controlled artefact registries.

A few exceptions, such as Cyrus [23], VIS [43], and Z-Inspection® [4], incorporate repeatable measurement processes and artefact tracking. In particular, Z-Inspection® embeds a structured feedback loop across its Set-Up, Assess, and Resolve phases, enabling transparency requirements to be revisited and refined in light of new evidence or system changes. While the involved stakeholders are typically domain experts or project partners rather than end-users, this lifecycle-aware design demonstrates how continuous evaluation can be operationalised in practice.

A number of studies also reported unintended consequences of transparency practices. For instance, Lee et al. [40] noted increased user anxiety when information was disclosed without sufficient context, while Fehr et al. [22] highlighted concerns from developers regarding the exposure of proprietary information. These findings illustrate the need to balance transparency with usability, commercial confidentiality, and cognitive burden, ensuring that lifecycle-integrated evaluations do not inadvertently reduce user trust or hinder adoption.

This theme directly addresses RQ3 (strengths, limitations, and gaps) by identifying the absence of standardised definitions and cross-domain benchmarks, and RQ4 (applicability in real-world AI contexts) by showing that without lifecycle integration, transparency evaluation risks becoming a compliance snapshot rather than an ongoing governance practice. The findings support the need for artefact-driven, metadata-based approaches that can be embedded into development workflows, enabling continuous, comparable, and stakeholder-relevant transparency evaluation.

VII. Discussion

This section synthesises the key findings of the review in relation to the four research questions. It reflects on the strengths and limitations of current transparency evaluation practices in artificial intelligence, highlights existing gaps, and considers implications for future research and practice.

A. RQ1: Existing Approaches to Transparency Evaluation

Many studies introduced formal instruments such as ALTAI [17], POLARIS [28], and RETENTION [26], which aim to structure transparency evaluation through checklists or composite scoring. Others, like Z-Inspection® [4], offer qualitative, scenario-based frameworks grounded in ethical deliberation and stakeholder reflection. While these approaches enhance contextual sensitivity, they do not offer quantitative scoring or automation features.

Collectively, they reflect a growing recognition of the need to move beyond abstract principles and towards operational methods. Despite this progress, a clear pattern emerged: the majority of studies adopt qualitative or semi-structured methods, with limited attention to scoring granularity, automation, or lifecycle integration. Checklist-based approaches remain dominant but are often used without weighting, maturity models, or replicable metrics. Few studies propose evaluation instruments that can be implemented at scale or embedded into technical workflows.

B. RQ2: How Key Transparency Concerns are Addressed

Thematic analysis indicates that transparency is predominantly conceptualised as a multi-dimensional construct encompassing traceability, explainability, and communication. While most studies address at least one of these pillars, only a limited number propose integrated frameworks that operationalise all three in a cohesive manner.

Traceability and explainability are the most frequently addressed dimensions, commonly through system documentation, model-level interpretation methods, or audit trails. In contrast, communication, especially toward non-technical users or impacted stakeholders, remains significantly underdeveloped across the literature.

Although many frameworks reference values such as trust, usability, or ethical alignment, only a minority explicitly incorporate stakeholder or end-user perspectives in the evaluation process. Notable exceptions include the FAT-CAT framework [40] and the transparency survey developed by Fehr et al. [22], both of which engage user feedback more directly. However, such examples remain rare.

This highlights a persistent gap between the normative emphasis on stakeholder-centred transparency and the current reality of system- or expert-centric evaluation practices. Most frameworks still prioritise internal validation or technical explainability over participatory or communicative transparency mechanisms.

C. RQ3: Strengths, Limitations, and Gaps

Several strengths can be observed in the reviewed literature. First, there is increasing convergence around the idea that transparency must be evaluable, not merely aspirational. Second, the inclusion of sector-specific indicators, particularly in healthcare, demonstrates the potential for contextual adaptation. Third, some studies have made significant strides

in mapping transparency to legal or ethical requirements. Nonetheless, persistent limitations were identified:

- Over-reliance on checklists and qualitative assessments, with limited support for scoring, automation, or benchmarking.
- Lack of standardised frameworks or maturity models for transparency.
- Limited stakeholder inclusion, especially in the evaluation of communication clarity or user understanding.
- Inconsistent treatment of lifecycle phases, with transparency often evaluated only at the design stage or after.
- Conceptual conflation between transparency, explainability, and related constructs.

These limitations mirror concerns raised in prior literature. For instance, Mittelstadt et al. [45] warned against ambiguity in transparency definitions, arguing that a lack of clarity undermines both the interpretability of AI systems and the accountability of their developers.

D. RQ4: Applicability in Real-World AI Contexts

Of the 28 studies reviewed, 23 (82%) [43], [30], [41], [31], [21], [4], [39], [22], [23], [44], [40], [35], [33], [27], [34], [36], [37], [28], [24], [25], [38], [26], [29] were applied in real-world contexts, suggesting a growing emphasis on empirical validation. These included pilot implementations, case studies, and evaluations of deployed systems, particularly in healthcare and cybersecurity domains. However, even among applied studies, evaluation methods often lacked formal mechanisms for compliance verification, stakeholder feedback loops, or documentation updates over time. Moreover, the strong concentration of studies in healthcare raises questions about generalisability. While medical AI offers a high-risk, high-regulation testbed, other domains, such as education, employment, or defence, are underexplored. This suggests that many existing frameworks may not yet be transferable to broader or less-regulated contexts.

E. Additional Observations and Future Trends

Emerging findings suggest that increased transparency can paradoxically raise user anxiety or cognitive burden, as demonstrated by Lee et al. [40]. Furthermore, several studies highlight the potential need for transparency toolkits, particularly for SMEs, to operationalise legal requirements in accessible ways. Concepts such as real-time AI dashboards ('AI cockpits') [33] or safe harbour mechanisms [31] also represent future directions worth exploring.

F. Synthesis and Implications for Future Research

Taken together, the findings highlight both progress and persistent gaps in the operationalisation of transparency in AI systems. While concrete evaluation methods are emerging, their diversity in scope, terminology, and measurement approaches underscores a lack of standardisation that complicates benchmarking across domains.

1) Fragmentation and standardisation challenges: Similar concepts, such as traceability or interpretability, are often operationalised differently, with artefacts ranging from checklists and surveys to explainability-based metrics. This fragmentation limits cross-domain comparability and reinforces the need for adaptable yet standardised schemes.

2) Lifecycle integration and sustainability: Most methods assess transparency only at discrete points, making it a compliance "snapshot" rather than an ongoing governance practice. Exceptions such as Cyrus and VIS point to lifecycle-aware approaches, but these remain isolated and not widely adopted.

3) Stakeholder perspectives and unintended consequences: Few studies systematically include user perspectives. Where they do, such as FAT-CAT [40] or Fehr et al. [22], results show both benefits and unintended consequences (e.g., increased anxiety when disclosure lacks context, or concerns about exposing proprietary information). This underscores the tension between usability, confidentiality, and cognitive burden.

4) Quality assessment insights: The quality assessment (QCA1-QCA5) provides additional perspective. While QCA1 (clarity of evaluation method) and QCA3 (discussion of strengths and limitations) scored highly (over 90%), QCA4 (comparison with other frameworks) was much weaker at 46.4%. As a result, transparency evaluation remains fragmented and difficult to contextualise, undermining comparability across sectors.

5) Bridging ethics and engineering: The three-pillar framing of traceability, explainability, and communication, first articulated by the AI HLEG [1], offers a conceptual anchor. Mapping these pillars to quality factors from software engineering (e.g., accessibility, usability, understandability) suggests a bridge between high-level ethical principles and measurable lifecycle practices.

Beyond these priorities, the review highlights an overarching limitation: the lack of artefact-driven methods capable of embedding transparency evaluation into day-to-day development workflows. Existing frameworks rarely provide mechanisms for capturing, standardising, and reusing transparency artefacts across system versions and contexts. This is a critical gap given the need for reproducibility, comparability, and regulatory accountability in real-world AI applications.

Addressing this limitation directly informed our subsequent work—a companion paper, which extends the Z-Inspection® process within the MANOLO context by proposing a Transparency Artefact Registry. TAR operationalises the insights of this SLR by offering a structured, metadata-based approach for cataloguing transparency artefacts, linking them to stakeholder requirements, and supporting continuous lifecycle evaluation. In doing so, it aims to provide researchers, practitioners, and regulators with a verifiable and repeatable mechanism for demonstrating transparency in line with European regulatory and ethical standards.

6) Design implications: The deficits identified map to an artefact-driven remediation. Missing standardisation calls

for a minimal metadata schema that normalises transparency artefacts across projects. Weak lifecycle integration requires a versioned registry that tracks artefacts across releases and updates. Informal regulatory alignment is addressed by explicit field-level mappings from artefacts to EU AI Act obligations (e.g., Article 13 on traceability and user information). Limited stakeholder inclusion is realised by tagging artefacts with stakeholder roles and providing communication-ready views. These elements motivate the TAR and align with Z-Inspection®.

Ultimately, transparency should not be treated as a static artefact or the outcome of a checklist, but as a dynamic, context-sensitive, and stakeholder-driven attribute that must be continuously embedded throughout the AI system development lifecycle.

VIII. Conclusion

This paper has presented an SLR of 28 peer-reviewed studies on transparency evaluation in artificial intelligence systems. The analysis revealed that while transparency is widely recognised as a requirement for trustworthy and accountable AI, current evaluation methods remain fragmented. Most approaches rely on qualitative checklists or scenario-based assessments, with limited support for scoring, standardisation, or lifecycle integration. Stakeholder involvement, particularly from end-users, is still rare. In particular, the communication dimension of transparency, meaning the disclosure and presentation of information about AI systems to users, regulators, and other stakeholders in accessible and understandable ways, remains underdeveloped compared to traceability and explainability.

Despite these limitations, the review identified encouraging trends. Several studies successfully link transparency evaluation to regulatory or ethical requirements, others adapt sector-specific indicators (notably in healthcare), and a minority demonstrate mechanisms for empirical validation in real-world systems. Together, these works mark a gradual shift from aspirational principles to operational practices.

At the same time, persistent gaps were observed: the lack of shared taxonomies and maturity models, insufficient mechanisms for continuous evaluation across the AI lifecycle, and the absence of artefact-driven methods for capturing and reusing transparency evidence. These findings highlight both the strengths of current approaches and the need for more systematic, repeatable, and scalable solutions.

The results of this review directly inform our subsequent work within the context of the MANOLO project, where we extend the Z-Inspection® Framework with a Transparency Artefact Registry (TAR). By cataloguing transparency artefacts through structured metadata and linking them to stakeholder requirements and regulatory obligations, TAR provides a practical means of addressing the limitations identified in this review. In this way, the SLR serves as both a state-of-the-art assessment and the conceptual foundation for developing artefact-based mechanisms that enable transparency to be continuously evaluated and demonstrably aligned with the EU AI Act.

In sum, transparency evaluation is moving from principle to practice but remains methodologically fragmented and weakly grounded across the AI lifecycle. Closing this gap requires standardised taxonomies, comparable metrics, and evidence-bearing artefacts embedded in development and compliance workflows.

Future research should therefore focus on refining such artefact-driven approaches, testing their applicability across domains beyond healthcare, and ensuring that stakeholder perspectives are integrated in a sustained and iterative manner. Only by embedding transparency throughout the lifecycle of AI systems can it become a verifiable attribute of trustworthy AI rather than a static aspiration.

Acknowledgment

This work is based on and extends the thesis by Karanxha [46], completed at Laurea University of Applied Sciences, Finland. This work was co-funded by the European Union under GA no. 101135782. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or CNECT. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy ai," European Commission, 2019, accessed 17 May 2025. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[2] "Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence (artificial intelligence act)," Official Journal of the European Union, L 168/1, 2024, accessed 17 May 2025. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

[3] "Manolo project," https://manolo-project.eu/, 2023, accessed: 17-Aug-2025.

[4] R. V. Zicari, J. Brusseau, S. N. Blomberg, P. J. Honkoop, J. Haverinen, C. Kuziemsky, J. Laaksolahti, A. Lang, P. Linko, S. Mahmood, and et al., "Z-inspection®: A process to assess trustworthy ai," *IEEE Transactions on Technology and Society*, vol. 2, no. 2, pp. 83–97, 2021. [Online]. Available: https://doi.org/10.1109/TTS.2021.3066209

[5] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai," Berkman Klein Center for Internet & Society, Harvard University, Research Publication 2020-1, 2020. [Online]. Available: https://dx.doi.org/10.2139/ssrn.3518482

[6] M. Ebers, V. R. S. Hoch, F. Rosenkranz, H. Ruschemeier, and B. Steinrötter, "The european commission's proposal for an artificial intelligence act—a critical assessment by members of the robotics and ai law society (rails)," *J*, vol. 4, no. 4, pp. 589–603, 2021. [Online]. Available: https://doi.org/10.3390/j4040043

[7] R. A. Agha, G. Mathew, R. Rashid, A. Kerwan, A. Al-Jabir, C. Sohrabi, T. Franchi, M. Nicola, M. Agha, and T. Group, "Transparency in the reporting of artificial intelligence – the titan guideline," *Premier Journal of Science*, vol. 10, p. 100082, 2025. [Online]. Available: https://doi.org/10.70389/PJS.100082

[8] S. Genovesi, M. Haimerl, I. Merget, S. M. Prange, O. Obert, S. Wolf, and J. Ziehn, "Evaluating dimensions of ai transparency: A comparative study of standards, guidelines, and the eu ai act," in *Symposium on Scaling AI Assessments (SAIA 2024)*, ser. OpenAccess Series in Informatics (OASIcs), vol. TBD. Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2025, pp. 10:1–10:17. [Online]. Available: https://doi.org/10.4230/OASIcs.SAIA.2024.10

[9] M. Busuioc, "Accountable artificial intelligence: Holding algorithms to account," *Public Administration Review*, vol. 81, no. 5, pp. 825–836, 2021, open Access. [Online]. Available: https://doi.org/10.1111/puar.13293

[10] C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, and E. Gomez, "The role of explainable ai in the context of the ai act," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2023, pp. 1139–1150. [Online]. Available: https://doi.org/10.1145/3593013.3594069

[11] J. C. S. d. P. Leite and C. Cappelli, "Software transparency," *Business & Information Systems Engineering*, vol. 2, no. 3, pp. 127–139, 2010.

[12] M. Hosseini, A. Shahri, K. Phalp, and R. Ali, "Engineering transparency requirements: A modelling and analysis framework," *Information Systems*, vol. 74, pp. 3–22, 2018, information Systems Engineering: selected papers from CAiSE 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306437916305282

[13] D. Spagnuelo, C. Bartolini, and G. Lenzini, "Qualifying and measuring transparency: A medical data system case study," *Computers & Security*, vol. 91, p. 101717, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016740481830823X

[14] P. Ofem, B. Isong, and F. Lugayizi, "On the concept of transparency: A systematic literature review," *IEEE Access*, vol. 10, pp. 89 887–89 914, 2022. [Online]. Available: https://doi.org/10.1109/ACCESS.2022.3200487

[15] B. Isong, P. Ofem, and F. Lugayizi, "Towards a framework for improving transparency in the software engineering process," in *Proceedings of the 2022 12th International Conference on Software Technology and Engineering (ICSTE)*, Osaka, Japan, 2022, pp. 19–28. [Online]. Available: https://doi.org/10.1109/ICSTE57415.2022.00011

[16] P. Ofem, B. Isong, and F. Lugayizi, "Metrics for evaluating and improving transparency in software engineering: An empirical study and improvement model," *SN Computer Science*, vol. 5, no. 1097, pp. 1–16, 2024.

[17] High-Level Expert Group on Artificial Intelligence, "The assessment list for trustworthy artificial intelligence (altai) for self-assessment," Report, Brussels, 2020, accessed 17 May 2025. [Online]. Available: https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence

[18] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006. [Online]. Available: https://doi.org/10.1191/1478088706qp063oa

[19] R. Rokhshad, M. Ducret, A. Chaurasia, T. Karteva, M. Radenkovic, J. Roganovic, M. Hamdan, H. Mohammad-Rahimi, J. Krois, P. Lahoud, and F. Schwendicke, "Ethical considerations on artificial intelligence in dentistry: A framework and checklist," *Journal of Dentistry*, 2023.

[20] M. Mora-Cantallops, S. Sánchez-Alonso, E. García-Barriocanal, and M.-A. Sicilia, "Traceability for trustworthy ai: A review of models and tools," *Big Data and Cognitive Computing*, vol. 5, no. 2, p. 20, 2021. [Online]. Available: https://doi.org/10.3390/bdcc5

[21] D. W. Joyce, A. Kormilitzin, K. A. Smith, A. Simons, I. Craddock, and A. Nevado-Holgado, "Explainable artificial intelligence for mental health through transparency and interpretability for understandability," *npj Digital Medicine*, vol. 6, p. 6, 2023. [Online]. Available: https://doi.org/10.1038/s41746-023-00751-9

[22] J. Fehr, G. Jaramillo-Gutierrez, L. Oala, M. I. Gröschel, M. Bierwirth, P. Balachandran, A. Werneck-Leite, and C. Lippert, "Piloting a survey-based assessment of transparency and trustworthiness with three medical ai tools," *Healthcare*, vol. 10, no. 10, p. 1923, 2022. [Online]. Available: https://doi.org/10.3390/healthcare10101923

[23] M. Basereh, A. Caputo, and R. Brennan, "Automatic transparency evaluation for open knowledge extraction systems," *Journal of Biomedical Semantics*, vol. 14, p. 12, 2023. [Online]. Available: https://doi.org/10.1186/s13326-023-00293-9

[24] V. Janev, M. Nenadović, D. Paunović, S. Vahdati, J. Li, M. H. Yousuf, J. Montanya, R. Theilen, J. Wittenstein, S. Tsurkan, and R. Huhle, "IntelliLung AI-DSS Trustworthiness Evaluation Framework," in *Proceedings of the 2024 32nd Telecommunications Forum (TELFOR)*. IEEE, 2024. [Online]. Available: https://doi.org/10.1109/TELFOR63250.2024.10819068

[25] L. Helmer, C. Martens, D. Wegener, M. Akila, D. Becker, and S. Abbas, "Towards trustworthy ai engineering: A case study on integrating an ai audit catalog into mlops processes," in *Proceedings of the 2nd International Workshop on Responsible AI Engineering (RAIE '24)*. New York, NY, USA: Association for Computing Machinery, 2024, pp. 1–7. [Online]. Available: https://doi.org/10.1145/3643691.3648584

[26] I. E. Nicolae, G. Danciu, C. Nanou, N. Koulierakis, and V. Danilatou, "Transparency metrics for artificial intelligence-driven applications in healthcare," in *Proceedings of the 13th Hellenic Conference on Artificial Intelligence (SETN '24)*. New York, NY, USA: Association for Computing Machinery, 2024, pp. 1–8. [Online]. Available: https://doi.org/10.1145/3688671.3688782

[27] D. M. F. Saldanha, C. N. Dias, and S. Guillaumon, "Transparency and accountability in digital public services: Learning from the brazilian cases," *Government Information Quarterly*, vol. 39, no. 2, p. 101680, 2022. [Online]. Available: https://doi.org/10.1016/j.giq.2022.101680

[28] M. T. Baldassarre, D. Gigante, M. Kalinowski, and A. Ragone, "POLARIS: A framework to guide the development of trustworthy AI systems," in *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering - Software Engineering for AI (CAIN '24)*. New York: Association for Computing Machinery, 2024. [Online]. Available: https://doi.org/10.1145/3644815.3644947

[29] P. Salehi, Y. Ba, N. Kim, A. Mosallanezhad, A. Pan, M. C. Cohen, Y. Wang, J. Zhao, S. Bhatti, J. Sung, E. Blasch, M. V. Mancenido, and E. K. Chiou, "Towards trustworthy ai-enabled decision support systems: Validation of the multisource ai scorecard table (mast)," *Journal of Artificial Intelligence Research*, 2024. [Online]. Available: https://doi.org/10.1613/jair.1.14990

[30] 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 *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, 2020, pp. 33–44.

[31] S. Yanisky-Ravid and S. K. Hallisey, ""Equality and Privacy by Design"": A New Model of Artificial Intelligence Data Transparency via Auditing, Certification, and Safe Harbor Regimes," *Fordham Urban Law Journal*, vol. 46, no. 2, pp. 428–491, 2019, accessed 17 May 2025. [Online]. Available: https://ir.lawnet.fordham.edu/ulj/vol46/iss2/5

[32] N. A. Sharma, R. R. Chand, Z. Buksh, A. B. M. S. Ali, A. Hanif, and A. Beheshti, "Explainable ai frameworks: Navigating the present challenges and unveiling innovative applications," *Algorithms*, vol. 17, no. 6, p. 227, 2024. [Online]. Available: https://doi.org/10.3390/a17060227

[33] L. Jantzen, M. Bottel, and R. Kempen, "How to achieve trust, satisfaction, and acceptance in the interaction with ai through an ai cockpit and a transparency interface? – a psychological framework," *Procedia Computer Science*, vol. 246, pp. 292–301, 2024. [Online]. Available: https://doi.org/10.1016/j.procs.2024.09.408

[34] A. Dubey, Z. Yang, and G. Hattab, "A nested model for ai design and validation," *iScience*, vol. 27, no. 9, p. 1110603, 2024, open Access. [Online]. Available: https://doi.org/10.1016/j.isci.2024.1110603

[35] K. Sokol, R. Santos-Rodriguez, and P. Flach, "Fat forensics: A python toolbox for algorithmic fairness, accountability and transparency," *Software Impacts*, vol. 14, p. 100406, 2022. [Online]. Available: https://doi.org/10.1016/j.simpa.2022.100406

[36] J. Haurogné, N. Basheer, and S. Islam, "Vulnerability detection using bert based llm model with transparency obligation practice towards trustworthy ai," *Machine Learning with Applications*, vol. 18, p. 100598, 2024. [Online]. Available: https://doi.org/10.1016/j.mlwa.2024.100598

[37] S. V. S. Kumar and H. K. Kondaveeti, "Towards transparency in ai: Explainable bird species image classification for ecological research," *Ecological Indicators*, vol. 169, p. 112886, 2024. [Online]. Available: https://doi.org/10.1016/j.ecolind.2024.112886

[38] A. Bansal, D. A. A. K, S. Gangadharan, R. R. R. R, M. M. Ismail, and N. K. A. Halaf, "Ethical considerations of AI implementation in business planning: Ensuring fairness and transparency," in *Proceedings of the 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, Chennai, India, 2023, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICSES60034.2023.10465461

[39] T. M. van Engers and D. M. de Vries, "Governmental transparency in the era of artificial intelligence," in *Frontiers in Artificial Intelligence and Applications*, vol. 322, 2019, pp. 33–42. [Online]. Available: https://doi.org/10.3233/FAIA190304

[40] C. Lee and K. Cha, "Fat-cat explainability and augmentation for an ai system: A case study on ai recruitment-system adoption," *International Journal of Human-Computer Studies*, vol. 171, p. 102976, 2023. [Online]. Available: https://doi.org/10.1016/j.ijhcs.2022.102976

[41] M. Vössing, N. Kühl, M. Lind, K. Främling, and G. Satzger, "Designing transparency for effective human-ai collaboration," *Information Systems Frontiers*, vol. 24, pp. 877–895, 2022.

[42] O. Adamyk, O. Chereshnyuk, B. Adamyk, and S. Rylieiev, "Trustworthy ai: A fuzzy-multiple method for evaluating ethical principles in ai regulations," in *Proceedings of the 13th International Conference on Advanced Computer Information Technologies (ACIT)*, Wrocław, Poland, 2023. [Online]. Available: https://doi.org/10.1109/ACIT58437.2023.10275505

[43] I. Barclay, H. Taylor, A. Preece, I. Taylor, D. Verma, and G. de Mel, "A framework for fostering transparency in shared artificial intelligence models by increasing visibility of contributions," *Concurrency and Computation: Practice and Experience*, vol. 33, no. e6129, 2021.

[44] M. Mylrea and N. Robinson, "Artificial intelligence (ai) trust framework and maturity model: Applying an entropy lens to improve security, privacy, and ethical ai," *Entropy*, vol. 25, no. 10, p. 1429, 2023. [Online]. Available: https://doi.org/10.3390/e25101429

[45] B. Mittelstadt, "Interpretability and transparency in artificial intelligence," in *The Oxford Handbook of Digital Ethics*, C. Véliz, Ed. Oxford University Press, 2021, online edition, accessed 17 May 2025. [Online]. Available: https://doi.org/10.1093/oxfordhb/9780198857815.013.20

[46] G. Karanxha, "Evaluating transparency in artificial intelligence systems: Adapting the z-inspection® framework for the manolo project," Bachelor's Thesis, Laurea University of Applied Sciences, Espoo, Finland, 2025.

Auditable Real-Time Cold-Chain Monitoring with IoT and Blockchain Anchoring

Mohamed DOUBIZ[1], Mouad BANANE[2], Abdelali ZAKRANI[3], Allae ERRAISSI[4]
Laboratory of Artificial Intelligence-Modeling and Computational Engineering, Hassan II University, Morocco[1,2,3]
Chouaib Doukkali University, El Jadida, Morocco[4]

***Abstract*—Safe vaccine storage hinges on continuous, trustworthy temperature supervision and evidence that records have not been altered. Yet many cold rooms still rely on fragmented logging tools that lack real-time alerts, end-to-end traceability, and audit-ready data. This paper presents a practical, low-cost architecture that integrates Internet of Things (IoT) sensing with blockchain anchoring to deliver real-time monitoring and visibility, reliable anomaly detection, and tamper-evident provenance for cold-chain storage. The design couples minute-level telemetry and dashboarding with alert debouncing/hysteresis to reduce false alarms, while anchoring hourly summaries and event alerts on-chain to create a verifiable trail without exposing raw data or incurring recurring fees during experimentation. A prototype in a vaccine cold-room scenario demonstrates that the approach is simple to deploy on commodity hardware, scales by adding rooms/sensors, and produces operator-friendly notifications alongside independently verifiable records. This combination of edge retention and cryptographic anchoring provides a pragmatic path for pharmacies, clinics, and warehouses to upgrade from basic loggers to transparent, audit-ready monitoring, bridging operational needs (alerts) and compliance needs (provenance) in one system.**

***Keywords*—Internet of Things; blockchain; cold chain; vaccine storage; real-time monitoring; tamper-evident; data provenance; traceability**

I. Introduction

In 2021, Morocco ordered over 60 million doses of COVID-19 vaccines from Sinopharm, Sinovac, and AstraZeneca [1]. Maintaining these vaccines between 2 °C and 8 °C remains a major logistical challenge, particularly in constrained environments. The COVID-19 pandemic made this clear at scale: unprecedented volumes moved in just-in-time flows among manufacturers, carriers, regional warehouses, vaccination centers, and mobile units, with each link responsible for continuous temperature monitoring and shared traceability. Some vaccines, such as Pfizer-BioNTech, required ultra-cold storage (−90 °C to −60 °C) upstream and, after thawing, allowed storage between 2 °C and 8 °C for up to 10 weeks [2], multiplying the thermal-regime transitions that had to be controlled and attested at every hand-off.

This complexity strengthened the demand for end-to-end traceability and tamper-evident records of thermal excursions; several studies therefore proposed IoT-plus-blockchain architectures to record, anchor, and share readings among stakeholders, with the promise of better combating counterfeiting and disputes [3] [4].

However, most existing proposals suffer from at least two limitations: high integration and operating costs (proprietary hardware, cloud services, transaction fees); and a lack of end-to-end evaluations showing how to reconcile continuous monitoring with tamper-evident anchoring at low cost.

This article addresses that gap by presenting and experimentally evaluating a low-cost IoT–blockchain integration architecture for a vaccine cold room. We ask: how can real-time supervision, immediate anomaly detection, and a verifiable proof of integrity be ensured while minimizing operational complexity and cost?

Our hypothesis is that the decoupling rates (fine-grained acquisition with parsimonious anchoring) combined with completeness control can achieve this objective.

Our contributions are as follows.

- An end-to-end prototype built from standard components and open-source building blocks: DS18B20 sensor → ESP8266 (NodeMCU) → MQTT (Mosquitto on Raspberry Pi 4).
- A Python collector (venv) logs measurements, triggers alerts if T < 2 °C or T > 8 °C (Telegram + on-chain event), and anchors an hourly statistical summary to the blockchain (number of samples, min/max/mean temperature, CSV hash).
- A frugal anchoring scheme: sampling is performed every minute (expected: 60 readings/hour), while on-chain anchoring is hourly via a smart contract (anchoring & alerts) on a local Ethereum node (geth, dev mode), eliminating transaction fees during trials. A completeness watchdog flags any hour with fewer than 60 readings (sensor, network, or logging loss).
- Alerting and Visualization Chain: Telegram immediately notifies threshold breaches and each hourly commit (count, min, max, mean, hash).
- A Node-RED dashboard provides live monitoring (gauge + time series) with set-point lines (2 °C/8 °C) and device online/offline status.

Organization Section II surveys related work on IoT–blockchain for the cold chain. Section III details the materials and methods (sensors, MQTT, Python collector, smart contract), including the alerting protocol, hourly anchoring, and the completeness watchdog. Section IV presents the results

and evaluation, including a comparison with representative approaches. Section V discusses limitations and extensions. Section VI concludes.

II. Literature Review

Digital vaccine supply chains (DVSCs) increasingly adopt sensing, connectivity, and data systems to monitor vaccines end-to-end and strengthen resilience [5]. Persistent quality and access gaps in resource-constrained contexts motivate pragmatic storage and distribution solutions that fit local constraints [6]. During COVID-19, large-scale social-media analytics further highlighted operational bottlenecks across logistics [7], [8]. Together, these streams justify lightweight, field-deployable *continuous* monitoring coupled with *tamper-evident* record-keeping as a practical research target.

Within healthcare, the Internet of Things (IoT) has progressed from pilots to domain applications; recent edited volumes synthesize IoT-blockchain convergence patterns, use cases, and architectural considerations [9], [10], [11]. For regulated cold-chain products such as vaccines, raw telemetry is insufficient without trustworthy context: integrity, provenance, and traceability are required to transform temperature streams into audit-ready evidence. Surveys of blockchain-based decentralized trust for IoT conclude that distributed ledgers offer tamper-resistant logs and programmable workflows (smart contracts), while emphasizing throughput, latency, and privacy trade-offs that guide design-to-cost choices [12].

From a supply-chain perspective, engineering-management work documents transparency and governance benefits of blockchain alongside adoption and interoperability barriers, issues that intensify when budgets and connectivity are limited [13]. Implementation studies show that permissioned frameworks (e.g., Hyperledger) can meet trust requirements at lower operational cost when value exchange is not the primary objective [14]. These insights motivate architectures that keep the blockchain surface minimal (hash anchoring / event notarization) while handling bulk telemetry off-chain—well aligned with a low-cost cold-room deployment.

Closer to medicine logistics, a blockchain-based traceability system for *cold-chain medicine logistics* demonstrates end-to-end traceability and automatic quality assessment with compact on-chain proofs and off-chain sensor data [15]. In the blood cold chain, analyses detail vulnerabilities (excursions, custody hand-offs, counterfeit risk) and argue for immutable, condition-aware traceability, directly analogous to vaccine storage and transport [16]. A vaccine-focused chapter surveys blockchain opportunities across the lifecycle (counterfeit mitigation, cold-chain compliance, smart-contract-based conformance) while noting scalability limits that favor minimalist, permissioned deployments [17].

Operations-research work on vaccine logistics consolidates objectives (wastage, service level, cost) and decision layers (facility, inventory, routing) and underlines the value of reliable, high-granularity data for calibration and evaluation [18]. In this sense, an IoT-blockchain stack is both a monitoring/compliance tool and a *data infrastructure*: continuous telemetry (IoT) anchored with immutable proofs (blockchain) yields decision-grade streams for control, alarms, and post-hoc audits. At the interface with public-facing technologies, scoping reviews of social media and digital tools in vaccination stress trustworthy records and transparent workflows for programs and communities, with blockchain among DVSC enablers [19].

Synthesis for the present work. Three design patterns consistently align with a cold-room experiment.

- Separation of concerns: retain high-rate temperature streams off-chain and anchor compact hashes/events on a permissioned ledger [12], [15], [16], [17];
- Programmable compliance: smart contracts for threshold alerts, excursion attestation, and custody milestones [12], [16], [17];
- Design-to-cost: lightweight nodes, batching, and cadence tuned to regulatory and network constraints typical of resource-limited settings [20], [6], [13], [14], [18], [19].

III. Materials and Methods

We built and operated a verifiable cold chain monitor for a pharmacy-grade chamber maintained between 2 °C and 8 °C. A waterproof DS18B20 probe, sampled by a NodeMCU ESP8266 every 60 s, publishes JSON over MQTT to a Raspberry Pi gateway connected via Wi-Fi (wlan0). On the Pi, a Python process ingests the stream, writes hour-aligned CSV files with milli-degree precision, and applies excursion detection with debounce and hysteresis. Immediate alerts and per-hour file-integrity digests (CSV SHA-256) are anchored on a local, feeless Geth developer chain; operators receive human-readable Telegram messages, and a Node-RED dashboard presents a live integer gauge plus a Temperature Evolution chart with fixed 2 °C and 8 °C reference lines. All parameters, file paths, and commands reported below mirror the deployed system, emphasizing reproducibility, auditability, and operational simplicity in resource-constrained settings (Fig.1).

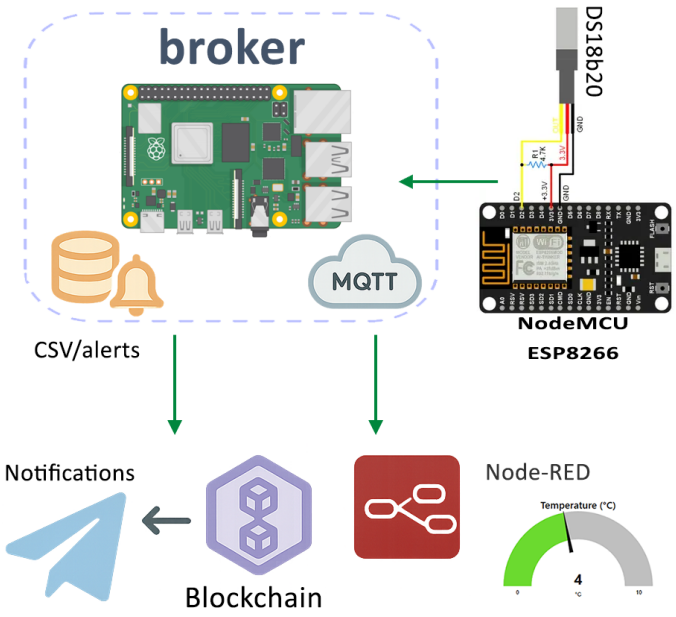


Fig. 1. Architecture: ESP8266 → MQTT → Pi monitor → CSV/alerts → on-chain → Telegram → Node-RED.

A. Hardware and Wiring (DS18B20 / ESP8266 / Raspberry Pi)

1) Cold room (process under control): The monitored environment whose air temperature must remain within 2 °C to 8 °C. It provides the real thermal dynamics (door openings, defrost cycles) against which the system is validated.

2) DS18B20 temperature probe (measurement): Digital 1-Wire thermometer (waterproof lead) that senses air temperature in the chamber. We placed the tip in the airflow, away from coils/doors to limit local bias. The resolution is fine enough for vaccine ranges; readings are consumed as is in this iteration (no correction applied).

4.7 KΩ pull-up (bus integrity). Keeps the 1-Wire data line idle-high and ensures reliable edges on long, thin leads.

3) NodeMCU ESP8266 V3 (sensor node & publisher):

a) Roles: powers the DS18B20 (3.3 V), polls it every 60 s, formats readings as JSON, and publishes via MQTT over Wi-Fi. It also maintains a retained liveness topic and a Last Will ("`offline`") for fault detection

b) What we store in each message: `temp_c` (float), `seq` (monotonic counter), and `ms` (`millis()`), which help the gateway spot gaps or reorders.

c) Why ESP8266: integrated 2.4 GHz Wi-Fi and very low BOM cost.

4) Raspberry Pi 4 (edge gateway & verifier):

- Roles: single-box aggregator on wlan0. It runs:
  - Mosquitto (MQTT broker) to receive ESP8266 messages,
  - our Python collector to append hour-aligned CSV, detect excursions (debounce + hysteresis), and compute SHA-256 digests,
  - Geth (dev mode) to anchor immediate alerts and hourly CSV integrity,
  - Telegram notifications, and
  - Node-RED dashboard (gauge + Temperature Evolution chart with 2 °C/ 8 °C reference lines).
- Networking: the Pi's wlan0 serves both the ESP8266 (same Wi-Fi) and Internet egress for Telegram.
- Wi-Fi access point/router (connectivity) Provides the WLAN where both ESP8266 and Pi's wlan0 are associated (DHCP + IP reachability). Internet access from this WLAN allows Telegram delivery.
- Power ESP8266 via stable 5 V USB (on-board 3.3 V regulator); Raspberry Pi via its PSU. A common ground exists between DS18B20 and ESP8266.
- Wiring we used
  - DS18B20 DATA → D2 (GPIO4) on ESP8266
  - DS18B20 VCC → 3.3 V, GND → GND
  - 4.7 KΩ resistor between DATA and 3.3 V.

B. Sensor Calibration

For this experiment we did not apply a numeric correction to the DS18B20. We performed a sanity check (ice-water bath $\approx 0\,^{\circ}\mathrm{C}$ and ambient $\approx 24\,^{\circ}\mathrm{C}$) and observed readings within the expected DS18B20 accuracy. Because the main objective was demonstrating the end-to-end chain (ingestion → detection → anchoring → notification), we accepted nominal accuracy and deferred formal calibration.

C. Firmware (Sampling, JSON, Topics, LWT)

File: `esp_ds18b20_mqtt.ino` (Arduino IDE).
Libraries: `ESP8266WiFi`, `PubSubClient`, `OneWire`, `DallasTemperature`.

- Sampling: every 60 s; the loop is non-blocking and calls `mqtt.loop()` continuously.
- JSON payload example:
  `{"temp_c": 5.500, "seq": 21, "ms": 1271041}`
  where `seq` increments per publish and `ms = millis()` assists drop/out-of-order detection
- Topics:
  - `coldchain/cold_room_1/status` (retained) → "`online`" at connect; LWT retained "`offline`".
  - `coldchain/cold_room_1/reading` (live) → readings (QoS 0)

D. Transport (MQTT): Broker, Auth, QoS, Tests

We ran Mosquitto on the Pi (listening on 0.0.0.0:1883 so the ESP8266 on Wi-Fi can connect).

- Auth: username/password (user `iotuser`, password stored in `/etc/mosquitto/passwd`).
- Config: `/etc / mosquitto / conf . d / coldchain.conf`
  with `allow_anonymous false`, persistence on, listener on `1883`.
- QoS: 0 to minimize latency; liveness comes from the retained `status`.
- Sanity tests we used repeatedly:

```
mosquitto_sub -h 127.0.0.1 -u iotuser -P
'********' -v -t 'coldchain/#'
mosquitto_pub -h 127.0.0.1 -u iotuser -P
'********' \  -t 'coldchain/cold_room_1/reading
' -m '{"temp_c":5.8,"seq":1,"ms":1000}'
```

E. Collector (CSV Windowing, Quality Policy, Timekeeping)

File: `/home/nuevomar/coldchain/monitor_from_mqtt_onchain.py` (Python 3.11 in a `.venv`). The process subscribes to `coldchain/+/reading` and `coldchain/+/status`, writes slot-based CSV archives, detects excursions, anchors integrity on-chain, and notifies Telegram. Runtime secrets (MQTT credentials, Telegram token, RPC URL) are loaded from `.env`.

Windowing → CSV: Readings are grouped into fixed slots of `SLOT_SECONDS`.

- Production profile: `SLOT_SECONDS = 3600` (one commit/hour).
- Lab profile: `SLOT_SECONDS = 60` (faster iteration).

CSV format: `iso_utc,epoch,temp_C,temp_mC`. We store milli-°C as integers to enable deterministic hashing.

Timekeeping: UTC timestamps from the Pi (NTP-synced). Slot boundaries aligned to the top of the hour in production.

Data quality: Outliers rejected if $|\Delta T/\Delta t| > 5\,^{\circ}\mathrm{C/min}$; missing/late samples recorded as `NaN` and excluded from averages, with `count` reflecting valid points. No smoothing (debounce/hysteresis covers noise).

Artifacts: hourly CSVs in `/var/log/coldchain/hourly/` and per-day alert logs in `/var/log/coldchain/alerts/`.

F. Alarm Logic (Band, Debounce, Hysteresis, Latency)

We enforced a $[2, 8]\,^{\circ}\mathrm{C}$ band with hysteresis and debounce:

- Thresholds: `LOW = 2.0 °C`, `HIGH = 8.0 °C`.
- Hysteresis: `HYST_MARGIN = 0.2 °C` ⇒ recovery band $[2.2, 7.8]\,^{\circ}\mathrm{C}$.
- Debounce: `DEBOUNCE_S = 180 s` (production); 30 s in lab.
- Events: `ALERT_START` after a sustained breach; `ALERT_END` after sustained recovery.
- Observed alert latency: $\approx 1.5-2$ min (lab profile, 60 s sampling / 30 s debounce) and $\approx 4-5$ min (production profile).

G. Blockchain Anchoring (Contract, Events, Schedule)

We used a local Geth developer chain (feeless, ephemeral) to anchor alerts and hourly CSV integrity.

Node startup used:

```
/usr/local/bin/geth --dev \
  --datadir /home/nuevomar/geth-dev \
  --http --http.addr 127.0.0.1 --http.port 8545 \
  --http.api eth,net,web3 --ipcdisable --dev.period 2
```

Contract: `ColdChainVax.sol` (deployed once). Artifacts persisted as `ColdChainVax.abi.json` and `ColdChainVax.address`.

Immediate anchoring (alerts): On `ALERT_START/END`, the collector called `submitReading(sensorId, epoch, temp_mC, kind)` with `sensorId = keccak256("cold_room_1")`, emitting an `Alert` event.

Periodic anchoring (per slot): The collector computed `(count, min, max, avg)` and the SHA-256 digest of the slot CSV, then emitted `HourCommitted(sensorId, slot, count, min_mC, max_mC, avg_mC, csvHash)`.

This yields tamper-evident on-chain proofs for alerts and per-hour CSV integrity while keeping full data off-chain.

H. Notifications (Telegram)

File: `/home/nuevomar/coldchain/notify.py` with `send_telegram(text)`; bot token and chat ID in `.env`.

Messages included:

```
[ALERT START] cold_room_1 2025-09-13T...Z T=...
 C (Above upper)
[CHAIN] submitReading tx: 0x...
[COMMIT] cold_room_1 slot=... count=... min=... max=...
avg=... hash=... tx=0x...
```

Using `wlan0` only, ensured outbound Telegram API access once Wi-Fi was up.

I. Visualization (Node-RED) + Reference-Line Configuration

We installed **Node-RED** with node-red-dashboard and built a single-tab dashboard:

- **Gauge** (°C): receives rounded integer values; Format `{{value | number:0}}`.
- **Temperature Evolution chart:** three series per measurement:
  - live temperature (`topic = cold_room_1`),
  - constant LOW line (`topic = LOW`, payload `2`),
  - constant HIGH line (`topic = HIGH`, payload `8`).
- The function node (two outputs) emits three messages to the chart and a rounded value to the gauge. Chart settings: Legend=ON, Use-one-color=OFF.

The flow is stored at `~/.node-red/flows_<hostname>.json`.

J. Security and Deployment (Secrets, Network, Persistence)

Secrets (`.env`) were set to file mode 0600 and not versioned; Telegram tokens were rotated as needed. MQTT required username/password; the broker listened on `0.0.0.0:1883` so the ESP8266 on the same Wi-Fi could reach it (TLS was not required on this isolated WLAN). Write access to `/var/log/coldchain/*` was restricted. We ran the collector as a single `systemd` unit (`coldchain-monitor-mqtt.service`) when needed to prevent multiple instances. CSV and alert logs persisted under

/var/log/coldchain; retention/backup can be handled with logrotate/rsync. (we preserved full logs during the experiment).

K. Software Versions and Reproducibility Commands

Environment we used As summarized in Table I, our deployment used commodity components and mainstream toolchains, which favors reproducibility.

TABLE I. SOFTWARE ENVIRONMENT USED IN THIS DEPLOYMENT

Layer	Component	Version / Notes
Hardware	Raspberry Pi	Pi 4, 64-bit OS
Firmware	ESP8266 Arduino core	3.x
Sensor libs	OneWire / DallasTemperature	2.x / 3.x
MQTT client	PubSubClient (ESP)	2.x
Messaging	Mosquitto	2.x (system pkg)
OS runtime	Python	3.11
Python pkgs	paho-mqtt / web3 / dotenv	1.6.x / 6.x / 1.x
Blockchain	Geth	v1.16.3 (-dev)
Smart contract	Solidity	0.8.20 (ABI/address persisted)
Node runtime	Node.js	v22.19.0
Edge app	Node-RED (+ dashboard)	3.x (+ node-red-dashboard 3.x)

Commands actually used:

```
# Python venv (+ deps)
cd /home/nuevomar/coldchain
python3 -m venv .venv && source .venv/bin/activate
pip install -U pip paho-mqtt web3 python-dotenv

# Start Geth dev (feeless, ephemeral)
mkdir -p /home/nuevomar/geth-dev
/usr/local/bin/geth --dev \
  --datadir /home/nuevomar/geth-dev \
  --http --http.addr 127.0.0.1 --http.port 8545 \
  --http.api eth,net,web3 --ipcdisable --dev.period 2

# Export runtime env (edit .env beforehand)
set -a; . ./.env; set +a
export RPC_URL=http://127.0.0.1:8545

# Run the monitor
python monitor_from_mqtt_onchain.py

# Check MQTT intake (during tests)
mosquitto_sub -h 127.0.0.1 -u iotuser -P '********' \
  -v -t 'coldchain/#'

# Inspect latest CSV & digest
LAST=$(ls -1t /var/log/coldchain/hourly/\
cold_room_1_*.csv | head -n1)
tail -n5 "$LAST"
sha256sum "$LAST" | cut -d' ' -f1
```

IV. RESULTS AND EVALUATION

A. Setup and Metrics

We sample once per minute (ESP8266→MQTT→Raspberry Pi), aggregate readings into fixed hour-aligned slots, and anchor both immediate alerts and per-slot integrity on chain. Two operating profiles were exercised: a *test profile* (debounce = 30 s) and a *production profile* (debounce = 180 s). We report alert latency (telemetry → alert emission), excursion statistics (count, total minutes out of band, median/P95 excursion duration), and integrity checks (on-chain digest vs. local CSV hash).

B. Network Observations and Mitigation (Wi-Fi → LAN)

During early trials on an isolated WLAN (MQTT QoS 0), we observed intermittent disconnections of the gateway interface (wlan0), visible as Last Will & Testament (LWT) toggles and short gaps in ingestion. Minor hiccups were absorbed by the collector's buffering, but multi-second drops inflated end-to-end alert latency and occasionally deferred Telegram delivery. To remove this confounder for the evaluation, we migrated the gateway to wired Ethernet (eth0); subsequent runs showed a stable 60 s cadence with no disconnects, while the anchoring logic and results remained unchanged.

C. Telemetry Ingestion Over Wi-Fi/MQTT

The ESP8266 (NodeMCU) published one DS18B20 measurement every 60 s over Wi-Fi to the Raspberry Pi Mosquitto broker (wlan0). The Python collector subscribed to coldchain/+/reading, persisted time-windowed CSV files, and computed per-slot statistics.

Minute-level telemetry and slot summary. In a representative one-hour window (slot 488279), we observe 60 samples at a stable 60 s cadence (median/P95 60 s/60 s). Temperatures range from 2.250 °C to 8.563 °C with a mean of 5.025 °C. Band compliance is 54 / 60 (90%) inside [2 °C, 8 °C]; 6 above 8 °C and 0 below 2 °C. The first out-of-range sample above 8 °C occurs at 2025-09-13T23:21:51Z; the last at 23:26:51Z (Fig. 2).

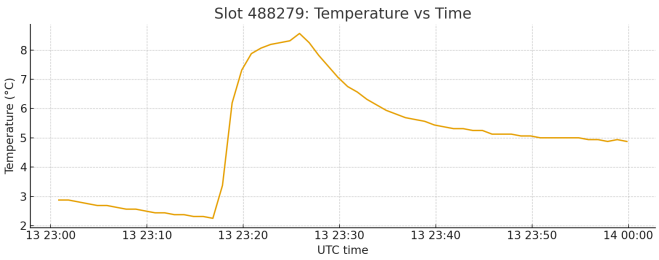


Fig. 2. Slot 488279 time series. The trace shows a brief excursion above 8 °C followed by recovery toward 5 °C.

For the representative slot 488279 (60 min window), we observe Table II:

TABLE II. SLOT 488279 SUMMARY

Rows	60
Sampling interval (median/P95)	60 s / 60 s
Min / Max / Avg	2.250 °C / 8.563 °C / 5.025 °C
Inside band [2 °C, 8 °C]	54 / 60 (90%)
Above 8 °C	6 / 60 (10%)
Below 2 °C	0 / 60 (0%)
CSV SHA-256	105c9c8...b0d49

D. Alarm Detection (2–8 °C Band; Debounce & Hysteresis)

Alerting behavior and latency. On 2025-09-13, the system detected 7 complete excursions (ALERT_START/END)

totaling 13 620 s out of range; median excursion 600 s, P95 6336 s. These timings align with the configured sampling and debounce: under the test profile (60 s sampling, 30 s debounce), end-to-end alert latency was ∼1.5–2 min; under production parameters, ∼4–5 min.

The collector enforced a 2 °C to 8 °C band with 0.2 °C hysteresis and a debounce window (test profile: 30 s; production: 180 s). A visual timeline of the `START` → `END` intervals appears in Fig. 3.

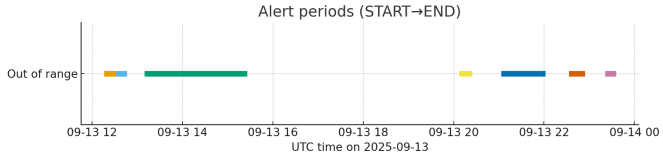


Fig. 3. Alert timeline on 2025-09-13 (START→END intervals).

E. On-Chain Anchoring (Devnet, Feeless)

We anchored both immediate alerts and slot integrity on a local Geth developer chain.

- **Alerts.** Each `ALERT_START/END` triggers `submitReading(sensorId, epoch, temp_mC, kind)`, emitting an `Alert` event; Telegram messages include the transaction hash (e.g., `[CHAIN] submitReading tx: 0xcbc8619c...`).
- **Hourly (slot) commits.** At the end of each slot we compute (count, min, max, avg) and the SHA-256 of the slot CSV and emit `HourCommitted(sensorId, slot, count, min_mC, max_mC, avg_mC, csvHash)`.

For slot 488279, the CSV file's digest exactly matches the on-chain `csvHash` reported in Telegram (`105c9c8...b0d49`), demonstrating reproducible integrity anchoring.

F. Human-in-the-Loop Notifications (Telegram)

The system delivered human-readable messages for `ALERT_START/END`, on-chain submissions, and per-slot commits. Once the Pi default route used `wlan0`, Telegram delivery was reliable. A representative excerpt is shown in Fig. 4, including an alert burst and two consecutive slot commits with their hashes and transaction IDs.

G. Visualization (Node-RED)

The dashboard (Fig. 5) presents an integer gauge and a Temperature Evolution Chart with continuous reference lines at 2 °C and 8 °C. Transient excursions and recovery dynamics are clearly visible; retained MQTT `status` reflects sensor liveness.

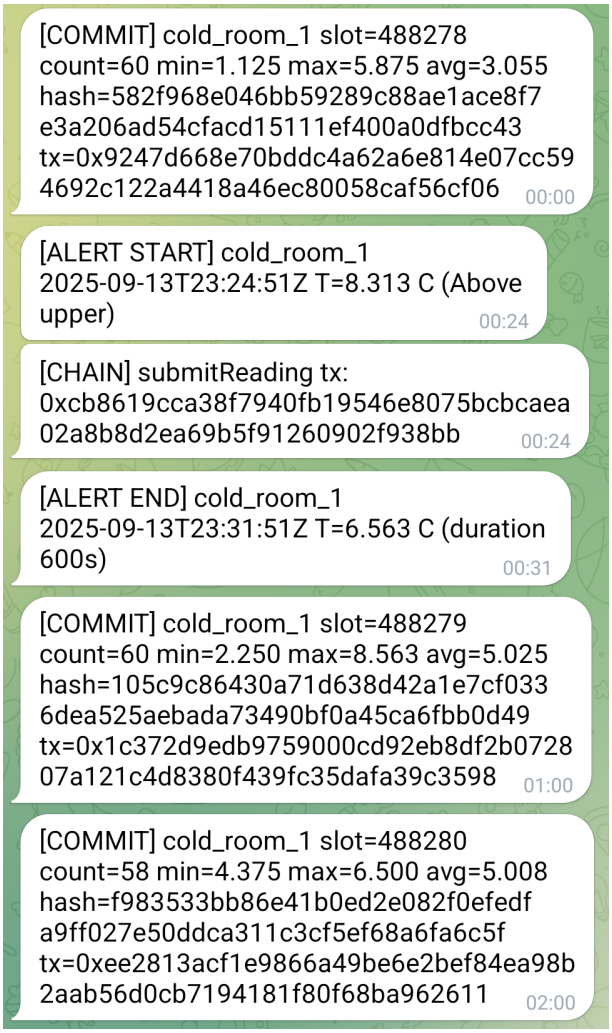


Fig. 4. Telegram excerpt showing an alert, the corresponding on-chain transaction, and per-slot commits with SHA-256 hashes.

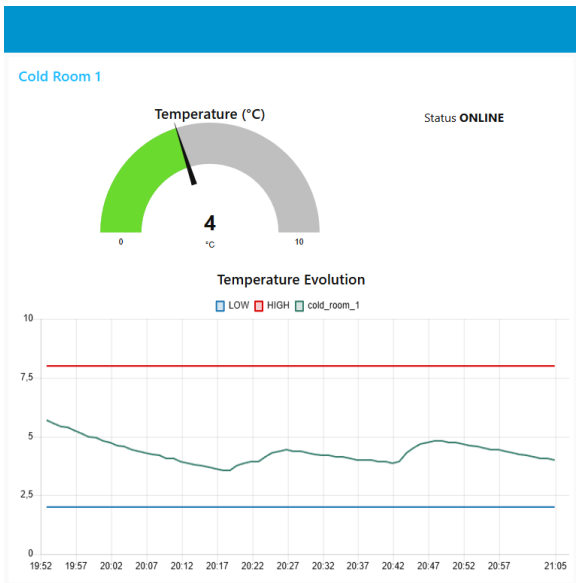


Fig. 5. Node-RED dashboard.

H. System Stability and Resource Use

On a Raspberry Pi 4, the stack (Mosquitto + Python collector + Node-RED + Geth dev) ran comfortably. Mosquitto and the collector exhibited low, bursty CPU at 1 msg/min; Geth's `-dev.period 2 s` caused modest periodic CPU spikes. After configuration stabilized, no restarts were required; a single systemd unit prevented concurrent collectors.

I. Integrity Verification Spot-Checks

We verified that:

- `sha256sum` of the slot CSV equals the on-chain `csvHash`;
- `eth_getTransactionReceipt` + ABI decoding reproduces `Alert` parameters;
- slot statistics `(count/min/max/avg)` recompute from the CSV and match the values posted on chain.

J. Slot Statistics

We partition the stream into fixed, hour-aligned slots ($\Delta_{\text{slot}} = 60\,\text{min}$). For slot 488279, we observe $N_s = 60$ records, median/P95 sampling interval 60 s/60 s, and temperatures in [2.250 °C, 8.563 °C] with a mean of 5.025 °C. Band compliance is 90% inside [2 °C, 8 °C], with 10% above 8 °C. The CSV's SHA-256 digest matches the on-chain `csvHash`, enabling independent verification of the stored file's integrity. Beyond per-slot completeness, we summarize anchoring regularity via inter-anchor intervals and report $p50/p90/p99$ when available.

K. Comparison with Representative Approaches

We contrast our decoupled design with two baselines used in the literature: continuous on-chain logging and batch-only notarization. The Table III is qualitative and highlights the trade-offs reviewers asked to see.

TABLE III. QUALITATIVE COMPARISON WITH REPRESENTATIVE BASELINES

Approach	Characteristics (from our deployment and standard baselines)
Continuous on-chain logging	Per-sample transactions; strong auditability but high transaction volume and on-chain data exposure; alerting often gated by confirmation time; cost scales with telemetry rate.
Batch-only notarization	Daily (or coarse) digests reduce cost but delay provenance; alerts typically derived post hoc; weaker near-real-time guarantees; low on-chain exposure.
This work (decoupled sparse anchoring)	**Real-time alerts from the live stream ;** hourly per-slot digests + event-level alerts provide verifiable provenance; raw data stay off chain (privacy); **cost bounded by slot cadence;** works on feeless devnet and is migration-ready for PoA/L2.

L. Cost Analysis and Total Cost of Ownership (TCO)

We analyse the costs for the implemented pipeline (DS18B20 → ESP8266 → MQTT/Mosquitto on Raspberry Pi → Python collector/CSV/alerts → hourly anchoring on local Geth (feeless) → Telegram/Node-RED), running at one reading per minute per room with hourly anchoring and immediate alert events. For concreteness, set the average powers to $P_{\text{Pi}} = 4\,\text{W}$ (light-load gateway) and $P_{\text{ESP}} = 0.15\,\text{W}$ (low-duty ESP8266 at 1 msg/min); with 24/7 operation the monthly gateway energy is $(0.004 + 0.00015\,N) \times 720 = 2.88 + 0.108N$ kWh for N rooms per gateway, i.e., per room $E_{\text{room}} = (2.88/N + 0.108)$ kWh/month. From the experiment logs (2025-09-13), one hourly CSV contains 60 samples in 2 671 B (min 2.250 °C, max 8.563 °C, mean 5.025 °C; 54/60 within $[2, 8]$ °C), extrapolating to ≈ 62.6 KB/day/room and ≈ 1.92 MB/month/room; the representative JSON payload `{"temp_c":5.500,"seq":21,"ms":1271041}` is $\approx$ 43 B, giving $\approx$ 1.86 MB/month/room of payload-only traffic, or $\approx$ 4.3–8.6 MB/month/room assuming a 100–200 B/message envelope (the broker is local; WAN usage is negligible). Anchoring is feeless in trials; on a PoA/L2 network the monthly on-chain volume per room is $720 + 2 \cdot \#\text{exc}$ transactions (hourly commits plus START/END), on 2025-09-13 there were 7 complete excursions (14 alert events), so the daily total would have been 24 hourly commits + 14 alerts = 38 transactions. This "alert-now, attest-later" scheme preserves near-real-time responsiveness while bounding OPEX by the hourly cadence and shared gateway resources (see Sections III–IV).

V. DISCUSSION

The experimental system delivered near–real-time visibility, reliable and latency-bounded alerts, and cryptographically verifiable provenance of stored data—using inexpensive hardware and a feeless local chain during testing. By combining edge CSV retention with per-slot hashing and event-level anchoring, the approach offers a pragmatic path to transparent, auditable cold-chain monitoring for vaccine storage while keeping operational complexity modest.

A. Contributions and Advantages

1) Real-time alerting without blockchain gating: In our deployment, alerts are produced directly from the live stream, achieving minute-scale responsiveness ($\approx 1.5\,\text{min}$ to $2\,\text{min}$) in the test profile and $\approx 4\,\text{min}$ to $5\,\text{min}$ in the production profile) rather than waiting for block confirmation. This preserves real-time operation while reserving the anchoring path for integrity and audit.

2) Tamper-evident provenance without exposing raw data: Event-level alert emissions and per-slot (hourly) CSV integrity digests (SHA-256) allow operators to recompute hashes and match on-chain values, yielding an auditable trail without placing raw telemetry on chain.

3) Operational simplicity on low-cost hardware: A single Raspberry Pi aggregates MQTT readings, applies debounce/hysteresis, writes hour-aligned CSV, and drives a Node-RED dashboard and Telegram notifications. The setup uses commodity components and documented versions/commands, emphasizing reproducibility.

4) Scalability and predictable cost: Anchoring at a slot/hour cadence bounds the number of commits—and therefore cost—while the topic-based design scales to multiple rooms. The same pattern can migrate from a feeless local

devnet to a private PoA or low-cost L2 without changing application logic.

5) Actionable operator visibility: The dashboard presents the averaged room series with fixed 2 °C and 8 °C reference lines, liveness indicators, and notifications, turning telemetry into decisions while retaining verifiable integrity of the underlying records.

B. Alignment with Objectives

1) Real-time visibility: With a 60 s sampling interval (ESP8266→MQTT→Raspberry Pi over Wi-Fi), the system provided continuous telemetry on a Node-RED dashboard (gauge + Temperature Evolution chart) and retained a minute-level historical trace. In a representative one-hour window (slot 488279) we observed 60 samples with 2.250 °C to 8.563 °C (mean 5.025 °C). Minute-level resolution is operationally adequate for pharmacy/clinic cold rooms, and the data volume remains tractable for edge storage.

2) Reliable, Interpretable alerting: The collector enforced a 2 °C to 8 °C band with 0.2 °C hysteresis and debounce. The behavior matches the design intent: short spikes did not trigger alerts, while sustained deviations did. End-to-end alert latency was ($\approx$ 1.5 min to 2 min) (test profile) and $\approx$ 4 min to 5 min (production), satisfying the operational need to react to sustained thermal risk rather than instantaneous noise.

3) Transparency and tamper-evidence: Two on-chain commitments were used: an immediate alert event (`Alert`) carrying the offending sample and timestamp, and a per-slot integrity event (`HourCommitted`) posting (*count, min, max, avg*) and the SHA-256 digest of the slot CSV. For slot 488279, the local `sha256sum` exactly matched the on-chain `csvHash`, enabling third-party verification that the plotted/analyzed file is bit-for-bit identical to the original at commit time.

4) Secure(er) retention: All raw measurements were kept off chain as CSV on the Pi, while only digests and minimal statistics were published on chain. MQTT authentication (username/password) and file permission hardening for secrets (`.env`) were applied. This separation reduces privacy/volume risk while preserving public verifiability of data integrity.

C. Reliability, Latency, and Human-in-the-Loop Operation

Although MQTT ran at QoS 0 on an isolated WLAN, reliability was high because the sampling rate is modest (1 msg/min) and the retained Last Will & Testament (LWT) topic exposed liveness ("online/offline"). Debounce and hysteresis suppressed flapping near thresholds. Telegram messaging closed the loop for operators: each `[ALERT START]` was paired with an on-chain submission (`[CHAIN] submitReading tx:...`), and each slot closure yielded a `[COMMIT] ...hash=...` line. From an operational standpoint, the latency budget is dominated by the sampling period and debounce; measured timings indicate the stack introduces minimal additional delay, which is appropriate for cold-chain supervision where cumulative exposure governs disposition decisions.

D. Auditability via Local, Feeless Anchoring

A low-friction audit path is available: an operator (or auditor) recomputes a slot CSV's SHA-256, retrieves the corresponding `HourCommitted` log, and verifies a match without accessing raw on-chain data beyond the digest and summary statistics. Using a local Geth developer network (`-dev`) removed transaction fees during experimentation and provided deterministic inclusion (typically $\sim$2–3 s with `-dev.period 2`). The anchoring pattern itself is chain-agnostic and can migrate to a private Proof-of-Authority network or a low-cost L2, preserving integrity guarantees while bounding cost through hourly (or daily) batching.

E. Security and Privacy Posture

The design keeps raw data off chain, publishes only hashes and aggregates, and authenticates MQTT clients. Secrets (MQTT password, Telegram token, RPC URL) are stored locally with restrictive permissions. While this posture is sufficient for a laboratory deployment, production environments should additionally enable TLS for MQTT and per-topic ACLs, rotate credentials, and consider a hardened key-management process for blockchain identities.

F. Generalizability and Scalability

The architecture scales by topic: additional rooms or probes publish to `coldchain/<room>/reading`, and the collector subscribes accordingly (wildcards). Edge compute load grows roughly linearly with sensor count and remains modest at minute-level sampling. On-chain costs scale with commit frequency, not raw telemetry size; therefore, batch anchoring (hourly/daily) keeps costs predictable. The same dashboard pattern (gauge + multi-series chart with reference lines) extends directly to multiple rooms.

G. Why the Design Works

The core advantage stems from *decoupling* sensing/alerting from anchoring. Alerts are produced directly from the live stream, so their latency is governed primarily by sampling and debounce rather than block confirmation. Hour-aligned slots provide a fixed cadence for integrity commitments, ensuring predictable anchoring overhead while keeping raw telemetry off chain. Room-level averaging (Eq. 1) concentrates operator attention on a single, representative series per room, reducing noise from probe-level idiosyncrasies yet preserving anomaly sensitivity through within-room spread and quorum checks.

H. Where it can Break (and How it Degrades)

1) Connectivity and buffering: With minute-level sampling, short Wi-Fi outages delay—but do not invalidate—alerts if the collector buffers; longer outages can push alert latency beyond the operational target. Hourly anchoring tolerates skipped anchors (later slots still authenticate prior CSVs), but prolonged loss creates gaps that must be explained operationally.

2) Sensing fidelity: Averaging can mask localized hot/cold spots if probes are poorly placed or if quorum degrades. The policy mitigates this via n_{valid} and the within-room spread (`std_C`); however, if $n_{\text{valid}} < 2$ the sample is marked missing, which reduces continuity for audit unless operators act on "quorum degraded" signals.

3) Chain and timing assumptions: Using a local developer chain removes fees and gives deterministic inclusion; migration to a private PoA or low-cost L2 preserves the pattern but changes absolute inclusion timings and operational procedures (RPC endpoints, keys, monitoring). The design remains chain-agnostic because it commits only digests and slot metadata.

I. Limitations and Threats to Validity

Ephemeral chain. The Geth developer chain is transient by design; for long-term evidence a persistent private chain or L2 must be used. The integrity mechanism remains unchanged.

- Single host and Wi-Fi only. The Raspberry Pi is a single point of failure; a UPS and hot-standby would increase availability. Wi-Fi loss would delay alerts; a wired or cellular fallback mitigates this risk.
- QoS 0 and no TLS (lab). Adequate for an isolated WLAN; regulated deployments should enable TLS and evaluate QoS based on loss/latency trade-offs.
- Temperature-only sensing. No door/humidity/compressor state was collected; adding these channels would improve root-cause attribution.
- Calibration not formalized. DS18B20 readings were sanity-checked but no traceable calibration curve was applied. A two-point correction (offset/slope) can be applied in firmware or at the collector without affecting anchoring.

J. Practical Implications

1) Parameterization: A production profile of 60 s sampling, 180 s debounce, and 0.2 °C hysteresis provided interpretable alerts while avoiding noise. Thresholds should match the product label (e.g., $+2$ to $+8$ °C for common vaccines).

2) Multi-sensor deployment with room-level averaging (recording policy): To improve accuracy and spatial representativeness while keeping the dataset compact, each cold room is equipped with ≥ 2 DS18B20 probes, but the *recorded and visualized data are the room-level average* rather than per-probe series. Let $T_i^{\mathrm{cal}}(t)$ denote the *calibrated* reading of probe i at time t (two-point offset/slope applied in firmware or at the collector). The stored value is

$$T_{\mathrm{room}}(t) \;=\; \frac{1}{N_t} \sum_{i \in \mathcal{V}(t)} T_i^{\mathrm{cal}}(t), \tag{1}$$

where $\mathcal{V}(t)$ is the set of *valid* probes at t (self-test passed, plausible range, not stale), and $N_t = |\mathcal{V}(t)|$. If fewer than a quorum (e.g., $N_t < 2$) are valid, the sample is marked missing.

CSV schema (recommended): store the averaged sample only, plus quality fields:

```
iso_utc, epoch, temp_C_mean,
temp_mC_mean, n_valid, std_C
```

As defined in Eq. (1), we average only valid probes at each timestamp. Here, `n_valid` is the number of valid probes contributing to Eq. (1) (i.e., `n_valid` $= N_t$), and `std_C` is the within-room standard deviation (a compact indicator of spatial spread without retaining per-probe traces).

3) Anchoring cadence (room-level only): Keep one on-chain commit per slot per room (e.g., hourly in production) with (count, min, max, avg) computed from the averaged series and the SHA-256 of the *room-level* CSV. Immediate alert events (`Alert`) are emitted only for room-level breaches. This preserves a tamper-evident trail while avoiding per-probe on-chain artifacts.

4) Multi-channel notifications: For operational resilience, deliver alerts in parallel:

- Telegram (already implemented) for rapid chat notifications,
- SMS (via provider API or GSM HAT) for reachability when data services are limited,
- E-mail (SMTP) for audit-friendly, searchable records.

Use a common message template (room ID, UTC time, averaged value, breach type, on-chain tx/hash), with delivery retries, rate-limiting (minimum interval between identical alerts), and escalation (e.g., notify a second contact after 10 min unresolved). Maintain daily liveness pings on all channels. In regulated environments, configure SMTP with SPF/DKIM/DMARC and manage provider API keys as encrypted secrets.

5) Operator workflow: Dashboards present the *single averaged room series* with continuous 2 °C/8 °C reference lines. Expose `n_valid` and `std_C` alongside the gauge to surface sensor health and spatial dispersion without storing individual probe traces. Provide cumulative minutes out of range and mean kinetic temperature (MKT) as secondary KPIs.

6) Security and migration: The averaging policy does not change security posture: secrets remain local; MQTT keeps authentication; raw samples stay off chain. Migrating anchoring from the local developer chain to a private PoA or low-cost L2 network, and enabling MQTT TLS + ACLs, are incremental steps that leave application logic and the averaging workflow unchanged.

VI. Conclusion

We presented a practical IoT–blockchain architecture for vaccine cold rooms that *decouples* real-time alerting from ledger anchoring. By producing alerts directly from the live stream and periodically committing hour-aligned SHA-256 digests of room-level records (plus event-level alerts), the system delivers minute-scale responsiveness, tamper-evident provenance, and predictable anchoring overhead—all on commodity hardware and without exposing raw telemetry on chain.

Our experiments confirmed this design: with 60 s sampling and simple debounce/hysteresis, the pipeline achieved minute-scale end-to-end alert latency while suppressing transient spikes, and on-chain `csvHash` values exactly matched local CSV digests, enabling independent verification of stored data. The resulting operator loop (dashboard + notifications) turns telemetry into timely, auditable actions, and the topic-based design scales to multiple rooms while bounding cost by slot cadence.

A. Main Contribution

We contribute a low-cost, reproducible pipeline that separates *detection* from *attestation*—"alert now, attest later"—achieving real-time supervision with verifiable integrity and controllable cost.

B. Future Work: Cost and Deployment Studies

Perform a total cost of ownership (TCO) analysis comparing a private PoA chain and a low-cost L2 under different anchoring cadences (per-alert only, hourly+alerts, daily+alerts). Instrument the gateway to measure energy (W/kWh) for the full stack (`Mosquitto` + collector + `Node-RED` + chain client) and compare against alternatives (e.g., Raspberry Pi 4 vs. Pi Zero 2 W vs. x86 NUC/mini-PC). Report KPIs such as cost/room/month, cost/commit, kWh/day, and CPU/RAM headroom; derive the break-even cadence (hourly vs. daily) that minimizes cost/audit proof while meeting the SLA for evidence availability.

REFERENCES

[1] Y. Ait Hamdan, F. El Amerany, J. Desbrieres, A. Aghrinane, H. Oudadesse, and M. Rhazi, "The evolution of the global covid-19 epidemic in morocco and understanding the different therapeutic approaches of chitosan in the control of the pandemic," *Polymer Bulletin*, vol. 80, no. 10, pp. 10 633–10 659, 2023.

[2] M. Wallace, J. P. Collins, H. Moline, I. D. Plumb, M. Godfrey, R. L. Morgan, D. Campos-Outcalt, S. E. Oliver, K. Dooling, and J. W. Gargano, "Effectiveness of pfizer-biontech covid-19 vaccine as evidence for policy action: A rapid systematic review and meta-analysis of non-randomized studies," *PLoS One*, vol. 17, no. 12, p. e0278624, 2022.

[3] U. J. Munasinghe and M. N. Halgamuge, "Supply chain traceability and counterfeit detection of covid-19 vaccines using novel blockchain-based vacledger system," *Expert Systems with Applications*, vol. 228, p. 120293, 2023.

[4] A. Musamih, R. Jayaraman, K. Salah, H. R. Hasan, I. Yaqoob, and Y. Al-Hammadi, "Blockchain-based solution for distribution and delivery of covid-19 vaccines," *Ieee Access*, vol. 9, pp. 71 372–71 387, 2021.

[5] W. Ran, Z. Li, Y. Xue, and D. He, "Literature review: Current trends and future prospects of digital vaccine supply chain support technology," *Human Vaccines & Immunotherapeutics*, vol. 21, no. 1, p. 2553454, 2025.

[6] G. Y. Scott and et al., "Enhancing vaccine quality and accessibility: Strategies for efficient storage and distribution in resource-constrained environment," *Outbreak Management and Response*, vol. 1, no. 1, 2025.

[7] A. Erraissi and M. Banane, "Machine learning model to predict the number of cases contaminated by covid-19," *IJCDS Journal*, vol. 10, pp. 991–1001, 10 2021.

[8] E. Cano-Marin, D. Ribeiro-Soriano, A. Mardani, and C. Blanco Gonzalez-Tejero, "Exploring the challenges of the covid-19 vaccine supply chain using social media analytics: A global perspective," *Sustainable Technology and Entrepreneurship*, vol. 2, no. 3, p. 100047, 2023.

[9] D. Mohamed, M. Banane, A. Zakrani, and A. Erraissi, *Integrating IoT for Enhanced Traceability, Safety, and Real-Time Control in Smart Farming Systems*, 12 2024, pp. 199–206.

[10] B. el Khalyly, A. Belangour, A. Erraissi, and M. Banane, "Devops and microservices based internet of things meta-model," *International Journal of Emerging Trends in Engineering Research*, vol. 8, 09 2020.

[11] S. K. Gupta, J. Rosak-Szyrocka, A. Mittal, S. K. Singh, and O. Hrybiuk, Eds., *Blockchain-Enabled Internet of Things Applications in Healthcare: Current Practices and Future Directions*. Singapore: Bentham Science, 2025, eISBN: 978-981-5305-21-0.

[12] Q.-u.-A. Arshad, W. Z. Khan, F. Azam, M. K. Khan, H. Yu, and Y. B. Zikria, "Blockchain-based decentralized trust management in iot: systems, requirements and challenges," *Complex & Intelligent Systems*, vol. 9, no. 6, pp. 6155–6176, 2023.

[13] H. A. Fattahi Bafghi, "Leveraging blockchain technology for sustainable, transparent and efficient supply chain management: An integrative exploration from an engineering management perspective," 2024.

[14] R. C. Aguilera, M. P. Ortiz, J. P. Ortiz, and E. V. Lozada, "Funding hyperledger blockchain dapp for covid-19 pandemic," *Fractals*, vol. 30, no. 6, p. 22501286, 2022.

[15] W. Zeng, Y. Wang, K. Liang, J. Li, and X. Niu, "Advancing emergency supplies management: A blockchain-based traceability system for cold-chain medicine logistics," *Advanced Theory and Simulations*, vol. 7, no. 4, p. 2300704, 2024.

[16] D. Asokan, J. Sunny, V. M. Pillai, and H. V. Nath, "Blockchain technology: a troubleshooter for blood cold chains," *Journal of Global Operations and Strategic Sourcing*, vol. 15, no. 3, pp. 316–344, 2022.

[17] T. Mawela, H. Smuts, F. Adebesin, M. Hattingh, and G. Maramba, "Opportunities for blockchain technologies in vaccine supply chain management," in *Bio-Inspired Computing: Proceedings of IBICA 2023, Volume 5: Real World Applications*, ser. Lecture Notes in Networks and Systems. Cham: Springer, 2025.

[18] J. Locke, M. Taghavi, B. Mansouri, and A. Saif, "Optimizing vaccine logistics: A taxonomy and narrative review," *INFOR: Information Systems and Operational Research*, 2025.

[19] S. Zang, X. Zhang, Y. Xing, J. Chen, L. Lin, and Z. Hou, "Applications of social media and digital technologies in vaccination: Scoping review," JMIR Preprints, 2023, preprint; final published version available in *Journal of Medical Internet Research*.

[20] A. Erraissi and A. Belangour, "A big data security layer meta-model proposition," *Advances in Science Technology and Engineering Systems Journal*, vol. 4, pp. 409–418, 10 2019.

Systematic Literature Review of Reactive Jamming Attacks Mitigation Techniques in Internet of Things Networks

Enos Letsoalo[1], Topside Mathonsi[2], Tshimangadzo Tshilongamulenzhe[3], Daniel du Duplesis[4]
Department of Information Technology, Tshwane University of Technology, Pretoria, South Africa[1, 2, 3]
Department of Computer Science, University of Johannesburg, Johannesburg, South Africa[4]

***Abstract*—Internet of Things (IoT) networks have become a prevalently exploited research area in academia and industry. IoT networks benefit from a variety of applications, including smart cities, smart homes, intelligent transportation, smart agriculture, monitoring, surveillance, etc. The security challenges associated with IoT networks have been broadly studied in the literature. This systematic literature review (SLR) is aimed at reviewing the existing research studies on IoT networks' reactive jamming attacks, challenges, and mitigation. This SLR examined the research studies published between 2019 and 2024 within the popular electronic digital libraries. We selected 45 papers after a rigorous screening of published works to answer the proposed research questions. The outcomes of this SLR reported three major IoT network performance issues. The results showed that the existing mitigation methods are categorized as machine learning based, deception-based, statistical-based, radio frequency-based, game theory-based, and encryption-based. The results show that most methods can detect reactive jamming attacks with accuracy. However, those methods still require additional infrastructure, encryption systems, and lead to prolonged training delays due to large datasets, resulting in computational overhead and transmission delays. Furthermore, the methods are unable to provide a better defense response to reactive jamming attacks. This is because the methods cannot adequately deal with the increased power consumption of IoT devices, cannot minimize transmission delays, and cannot improve the packet delivery ratio. As a result, reactive jamming attacks continue to be prevalent in IoT networks.**

***Keywords*—IoT networks; reactive jamming attacks; mitigation methods; systematic literature review; electronic digital libraries**

I. Introduction

Internet of Things (IoT) [10] networks are prevalent in distinct application domains that include smart cities, smart homes, intelligent transportation, agriculture, monitoring, etc. Applications of the IoT, such as smart agriculture [1], smart cities [2], and healthcare [3] aim to enhance the quality of life, health, and safety in both rural and urban communities. The demand for these applications is rapidly increasing, leading to a projected count of approximately 29 billion IoT devices by 2030 [4].

IoT networks are vulnerable to reactive jamming attacks (RJA). During reactive jamming, a malicious node sends high-power signals into active channels to disrupt legitimate communication through intentional interference [21]. The technologies such as Bluetooth low energy, RFID, and 802.15.4 have a few shortcomings, namely: they are restricted to communication with short distances, they provide low data rates, and provide low throughput. These technologies are vulnerable to jamming attacks. BLE is susceptible to a reactive narrow-band jammer. This type of jamming attack emits the jamming signal on a single BLE advertising channel when a jammer detects a frame transmission that must be attacked [36].

According to [38], RFID systems' components, such as tags and readers, are considered as single entities, and therefore, DoS attacks by signal jamming in the air interface affect both components. Jamming can result in communication being interrupted, which will make the reading process unavailable. IEEE 802.15.4 is the implementation of WSNs. Jamming of 802.15.4 affects the quality-of-service factors, namely the link quality of the channel, the delivery and delay of packets, and the energy consumption [37]. The widespread use of 802.11ah (Wi-Fi) devices has led to ubiquitous access of information. Jamming of 802.11ah networks can expose the weaknesses of the devices, namely, limited battery life and low processing power [31], [5]. The performance of power-constrained IoT devices is directly impacted by inefficient utilization of their resources.

RJA poses a cybersecurity threat for domains such as business and economy. In industrial intelligent systems, disruption due to RJA can result in financial loss. Smart agricultural systems can be left unmonitored due to communication disruptions. In smart transportation, if the systems are disrupted by RJA, passengers and goods are undelivered, which can lead to the unavailability of services. In real application systems, such as monitoring and surveillance, RJA can disrupt real-time communication.

Prior to this study, there were few survey studies that were done on jamming attacks in wireless networks. The authors in [23] surveyed jamming attacks across all types of wireless networks. The authors in [39] surveyed jamming and anti-jamming techniques in wireless networks. The authors in [40] surveyed jamming attacks in wireless networks. The authors in [41] surveyed jamming mitigation in cognitive radio networks. Unlike previous studies, this study surveys RJA detection and prevention techniques in IoT networks. Furthermore, this research study identified the research gaps left by existing mitigation techniques and proposes a smart channel hopping model to address those gaps.

This research study presents a comprehensive survey on reactive jamming attacks (RJAs) mitigation strategies in IoT networks. This research study will systematically review the papers that are available in the literature relating to challenges posed by jamming attacks, detection of jamming attacks, and mitigating jamming attacks in IoT networks. The research studies that are analyzed are those published in the last five years, i.e., from 2019 to 2024.

The objectives are to offer readers with a holistic knowledge landscape of existing RJAs mitigation techniques. This research also focuses on the performance evaluation of the IoT devices under RJAs. Furthermore, the strengths and weaknesses of the existing RJAs mitigation techniques are presented. The main findings of this research study are that the existing RJAs mitigation techniques can be categorized into six categories. We were able to answer the questions in Section III.

Unlike the previous studies, this study reviews RJAs mitigation techniques in IoT networks. It further categorizes the existing techniques of mitigating RJAs into: Machine learning (ML)-based, deception-based, statistics-based, radio frequency-based, game theory-based, and encryption-based. In addition, this study evaluates the strengths and weaknesses of the techniques using the metrics, namely: power consumption of IoT devices, transmission delays, and packet delivery ratio. The contributions of the study are as follows:

- Firstly, we conduct a survey of existing techniques for mitigating RJAs in IoT networks.
- Secondly, we discuss existing methods for detecting and preventing reactive jamming attacks. We tabulate the results and outline the strengths and weaknesses of the techniques.
- Thirdly, we briefly introduce the smart channel hopping model, which is proposed to address the gaps that were not addressed by the existing mitigation techniques.
- Finally, we provide the conclusion and the future direction of the research.

The rest of the study is organized as follows: Section II discusses research methods, Section III discusses results and discussion, Section IV introduces a technical overview of the smart channel hopping model, and Section V presents the conclusion and future research directions.

II. Research Method

A systematic literature review (SLR) as a research study method was used to assess and interpret the research topics available in the literature. Systematic literature analysis is another choice for SLR. Kitchenham guidelines are more specific for performing this SLR. The SLR procedure for this research study involves three subsections, namely: research questions, search strategy, and studies' inclusion and exclusion criteria.

A. Research Questions

The main objective of this research study is to examine the existing mitigation methods against reactive jamming attacks in IoT networks. This research also focuses on performance evaluation of mitigation strategies from significant existing research studies. To achieve the aims and objectives of this SLR, four research questions were formulated as follows:

- RQ1: What are the existing reactive jamming attacks mitigation methods?
- RQ2: What are the strengths of existing mitigation methods?
- RQ3: What are the limitations of the existing methods?
- RQ4: What are the research gaps left unaddressed by the existing methods?

B. Search Strategy

This section focuses on how the search criteria were conducted in terms of the use of search keywords, electronic sources, reference management tools, and a search strategy. Each process is described in the subsections that will follow:

1) Search keywords: The developed research questions were used to derive the search keywords and strings. We have also included synonyms and alternatives. We took synonym keywords from the relevant literature on reactive jamming attacks in IoT network topics. During the search process, the following keywords were used:

``reactive jamming attacks mitigation in IoT networks", ``reactive jamming attacks ",

``reactive jamming attacks detecting and prevention methods in IoT networks",

``mitigating reactive jamming attacks in IoT networks", ``jamming attacks in IoT networks", "jamming attacks mitigation".

2) Electronic sources: The popular electronic digital libraries were explored to extract the most relevant conference papers and journal articles related to jamming attacks mitigation. Table I shows the list of common digital libraries that have been used to search for research papers and articles.

TABLE I. List of Digital Libraries

Source	URL
IEEE Xplore	www.ieeexplore.org
Research Gate	www.researchgate.net
ACM	www.acm.com
Google Scholars	www.scholar.google.com
Science Direct	www.sciencedirect.com

These digital libraries are the major sources of publications on areas of computer network field.

3) Reference management: The search keywords and their alternative used in above search strategy yielded many studies from the above-mentioned electronic sources. There was no electronic or automatic tool that was used to manage the reference. The process of reference management was done manually.

4) Search process: We launched a search process on the digital libraries to retrieve the relevant literature from journal articles, conference papers, and books. This search process resulted in many studies. This search process was followed by a technique that was applied for the selection process to filter out the irrelevant papers.

C. Criteria for Inclusion and Exclusion of Studies

The criteria for including and excluding the studies is presented in Table II.

TABLE II. INCLUSION AND EXCLUSION CRITERIA

Inclusion criteria	Exclusion criteria
Research studies that discuss reactive jamming attacks mitigation in IoT networks	Research studies that are published before 2019
Research studies that discuss reactive jamming attacks detection and prevention in IoT networks	Research studies that discuss reactive jamming attacks mitigation in other types of networks than IoT network
Research studies that discuss reactive jamming attacks in IoT networks	Research studies that discuss reactive jamming attacks detection and prevention in other types of networks than IoT network
Research studies that discuss anti-jamming in IoT networks	Research studies that discuss reactive jamming attacks anti-jamming in other types of networks than IoT network
Research studies that discuss reactive jamming attacks mitigation in access and sensor IoT networks	Research studies that discuss reactive jamming attacks mitigation in IoT backhaul networks

Fig. 1 shows the filtering process that was executed after the search criteria were applied in the digital electronic libraries above.

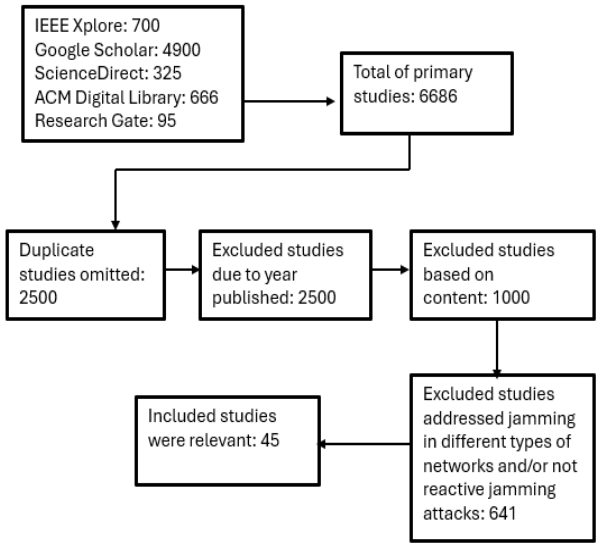


Fig. 1. Selection of studies process.

Fig. 2 presents the distribution of the selected papers by year. We note that 29% of the studies have been published in 2022 and then followed by 24% of the studies published in 2023.

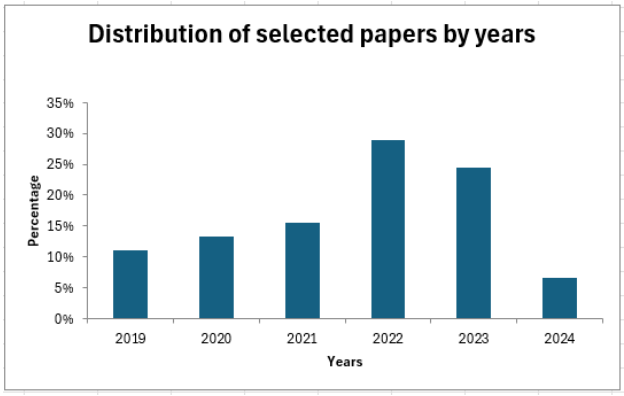


Fig. 2. Distribution of selected papers by years.

III. RESULTS AND DISCUSSIONS

In this section of our SLR, we analyze the most important solutions proposed, including an overview and comparison of categories of existing mitigation methods. Furthermore, analysis of the methods based on the metrics, namely strengths, limitations, the ability to respond to power consumption of IoT devices, transmission delays, and packet delivery ratio.

A. Categories of Mitigation Methods

In this section, the existing mitigation methods against reactive jamming attacks in IoT networks are presented. The techniques are categorized into six categories, namely: Machine learning (ML)-based, deception-based, statistics-based, radio frequency-based, game theory-based, and encryption-based.

1) ML-based methods: ML-based methods are those that employ machine learning techniques to detect and defend against RJAs. These methods utilize machine learning algorithms such as supervised learning. The pre-existing datasets are required for training the agents. They assume that the jammer does not change the jamming strategy, i.e., if the jammer changes the strategy, these methods may falter.

2) Deception-based methods: Deception-based methods use dummy frames or signals to lure the jammer into jamming the fake packets. Some methods use dedicated channels for deceiving the jammer. The transmitter requires to send fake signals prior to sending the real frames.

3) Statistics-based methods: Statistics-based methods use information extracted from the network to detect RJAs. The information may include received signal strength, packet delivery ratio, lack of acknowledgements, and packet error ratio. They lack defense response mechanisms.

4) Frequency-based methods: Frequency-based methods use frequency spectrum techniques for detecting and defending against RJAs. Techniques include patrolling the frequency channel through sensor nodes and using multiple antennas for signal backscattering during jamming attacks.

5) Game theory-based: Game theory-based methods are based on the concept of two players, namely the jammer and the legitimate node, who adjust power levels to defeat each order. Most methods use Q-learning algorithms to obtain the best policy for allocating optimal power to IoT devices to defeat

the jammer. The Q-learning algorithm often takes time to converge when the state space is too large.

6) Encryption-based methods: Encryption-based methods require encryption algorithms to be utilized to defeat reactive jamming attacks. IoT devices are constrained in terms of energy and computation. Therefore, it may not be feasible to implement encryption-based methods.

These mitigation methods are compared in Table III. The metrics that are used are computational overhead, defense response, and resource efficiency. Computational overhead assesses whether the method can be implemented with minimal strain on the energy and computational capacity of IoT devices. Defense response assesses whether the method can effectively mitigate the effects of reactive jamming attacks. Resource efficiency assesses if the method requires additional infrastructure for it to be implemented.

TABLE III. COMPARISON OF METHODS

Method	Computational Overhead	Defense response	Resource efficiency
ML-based	Medium. During agent training	Yes, when jammer not changing strategy.	Efficient
Deception-based	Medium. Frames sent for deception	Yes	Efficient
Statistics-based	Less	No	Efficient
Frequency-based	Medium. Requires sensor nodes and antennas.	Yes	Not Efficient
Game theory-based	Less	Yes, only after convergence	Efficient
Encryption-based	More. It requires encryption algorithms.	Yes	Not Efficient

B. Analysis of Mitigation Methods

This section presents the analysis of mitigation methods based on the metrics, namely strengths, limitations, the ability to respond to power consumption of IoT devices, transmission delays, and packet delivery ratio. Table III provides an analysis of the above-mentioned metrics. In addition, Table IV answers all the research questions raised in Section II A.

TABLE IV. STRENGTHS AND LIMITATIONS

Category	Author	Strength	Limitation	Power consumption	Packet delivery ratio	Transmission delays
Machine learning	Reynvoet et al. (2022)	High accuracy detection	Existing dataset reliance and training delays	Poor consumption due to overhead	Not considered	Increase due to overhead
	Upadyaya et al. (2019)	Detection accuracy	Introduce infrastructure and computational overhead	Poor during data collection and learning	Not considered	Increase due to overhead
(ML)-based	(Zahra et al., 2023)	95% detection accuracy	Training delays and centralized detection centre	Poor during data collection and learning collection and learning	Not considered	Increase due to algorithms complexity
	Lee et al. (2023)	High detection accuracy	Encryption leads to computational overhead	Improves by battery-drainage attack	Improves when users are less	Increases due to encryption
	Testi et al. (2023)	High detection accuracy	No defence response	Not considered	Not considered	Not considered
	Hussain et al. (2022)	86% detection accuracy	No defence response	Not considered	Not considered	Not considered
Deception-based	Thiha et al. (2019)	High detection accuracy	Overhead of sending fake frames	Poor consumption due to sending dummy frames	Improves when packet error ratio drops	Increases due to sending dummy frames
	Liu et al. (2021)	Increase throughput by 50%	Overhead of sending fake frames	Poor consumption due to sending decoy signals	Improves when packet error ratio drops	Increases due to sending decoy signals
	Pourranjbar et al. (2021)	Can efficiently deceive the jammer	Introduce additional channel and overhead of sending fake signals	Poor due to transmission into fake channel	Affects by delays	Increases due to transmission into fake channel
	Nan et al. (2020)	Can efficiently deceive the jammer	Introduce additional channel and overhead of sending fake signals	Improves when system has enough power budget	Not considered	Increases due to transmission into transmitter-receiver pair channel
	Hoang et al. 2020	Low packets drop ratio	Overhead of sending	Poor consumption	Affects by delays	Increases by sending

Category	Author	Strength	Limitation	Power consumption	Packet delivery ratio	Transmission delays
			fake frames			fake frames
	Huynh et al. 2021	Can efficiently deceive the jammer	Energy harvesting and backscattering only possible before channel jamming	Improves by energy harvesting techniue	Improves by backscattering	Increases during energy harvesting process
Statistics-based	Zhang et at. (2022)	High detection accuracy	No defence response	Not considered	Not considered	Not considered
	Singh et al. (2022)	PDR improved by 22%	No defence response	Not considered	Improves by 22%	Not considered
	Fadele et al. (2019)	High detection accuracy	High rates or bit errors and packet loss	Shows 3% power consumption improvement	Improved by 10%	Improved by 38%
	Abdollahi et al. 2023	High detection accuracy	Overhead is caused by states recording and reporting	Recording and reporting will drain the battery	Not considered	Not considered
Radio frequency-based	Arcangeloni et al, (2023)	99% detection accuracy	Introduces additional infrastructure and computational overhead	Poor due to computational overhead	Affects by computational overhead	Increases due to computational overhead
	Huynh et al. (2020)	Throughput can increase with more antennas	Requires more antennas	Poor due to power required by multiple antennas	Improves with addition of more antennas	Increases due backscattering
	Oukas et al. (2023)	99% detection accuracy	Energy harvesting will lead to low PDR	Improves by energy harvesting system	Not considered	Increases due many transition states
	Ali et al. 2023	Detection accuracy	Accuracy is affected by	Not considered	Affects by delays	Increases during NIC data
			NIC's vendor dependency			extraction phase
Game theory-based	Chkirbene et al. (2023)	Optimal power allocation can alleviate jamming attacks	Requires that the jammer does not rapidly change jamming strategy	Not considered	Improves after convergence	Convergence takes longer resulting in delays
	Gouissem et al. (2020)	Optimal power allocation can alleviate jamming attacks	Requires that the jammer does not rapidly change jamming strategy	Not considered	Improves after convergence	Convergence takes longer resulting in delays
	Gouissem et al. (2023)	Optimal power allocation can alleviate jamming attacks	Requires that the jammer does not rapidly change jamming strategy	Not considered	Improves after convergence	Converges faster but assumes jammer never change strategy.
Encryption-based	Navas et al. 2021	Resilient to jamming attacks	Introduces computational overhead dur to encryption	Power consumption due to computation	Affects by transmission delays	Increases due to computation

C. Discussion and Analysis

The ML-based studies proposed by [18], [25], [30], show that the reactive jamming attacks can be detected with high accuracy. However, these methods require that additional infrastructure and/or encryption systems be deployed as defense response mechanisms to the jamming attacks. This will lead to computational overhead which will result in transmission delays and low packer delivery ratio. The studies proposed by [15], [28], [32] indicate that the reactive jamming attacks can be detected with high accuracy. However, the proposed methods lack the defense response systems. This is because there is no mechanism for alleviating reactive jamming attacks when the attacks are detected. In addition, although prior knowledge of an un-jammed signal is unnecessary, ML-based jamming detection has difficulty in obtaining a large amount of training data, and the generalization ability of such

model is usually limited. Moreover, pre-existing datasets have a disadvantage that if the jammer rapidly changes the jamming strategy, they falter.

The deception-based studies proposed by [8], [14], [19], [20], [24], [29] show that sending fake frames into the normal channel or fake channel can deceive the reactive jammer into jamming fake frames. However, this will not be efficient in an environment where the sender must send many frames. This is because the IoT devices will spend more time sending fake frames between transmissions of real frames. This will lead to increased power consumption of power-constrained IoT devices, increased transmission delays of real frames, and a poor packet delivery ratio.

The methods that are proposed by [12], [27] show that reactive jamming attacks [26] can be detected with accuracy. The results show that packet delivery ratio can also be improved. In [27], the authors do not address the issue of power consumption and transmission delays. A study by [12] shows that transmission delays can be slightly improved under low traffic. Power consumption will increase when traffic increases due to congestion and retransmission. The method by [6], [33] shows that reactive jamming attacks can be detected with accuracy. However, the methods do not provide response mechanisms. The study by [6] introduces computational overhead. This is because IoT devices are required to regularly record channel states and send reports to the central node. This will quickly drain the IoT device battery. As a result, the issues of power consumption, packet delivery ratio, and transmission delays were not considered.

The methods proposed by [11], [16], [17] show that reactive jamming attacks can be detected with accuracy. However, the methods introduce additional infrastructure and computational overhead. This will impact transmission delays and packet delivery ratio. The proposed study by [22] achieves better detection accuracy and a response model. But the fact that it consists of many transition states brings several drawbacks, namely transmission delays, high power consumption and reduced packet delivery ratio. The study by [7] shows detection accuracy. However, the process of extracting information from the network interface card will introduce delays. In addition, since the NICs are manufactured by different vendors, this may affect the accuracy of method.

The methods proposed by [13], [34], [35] show that optimal power allocation to IoT devices can drastically reduce jamming attacks. However, these methods require the nodes to adjust or increase their transmission power levels to defeat reactive jammer. In addition, if the jammer has a higher power level, then jamming will not be controlled. Increasing power to the maximum levels will quickly drain the battery of the nodes. If the jammer can rapidly change the jamming strategy this method will fail. Furthermore, the prerequisite of network convergence for obtaining better defense policy may cause delays and affect throughput.

The encryption-based study proposed by [9] shows that the method can be resilient to jamming attacks. The strategy randomizes spread sequences used by the nodes in DSSS system. Every pair of communicating nodes will have a unique pairwise spreading sequence, only known by them. However, the study introduces a lot of computational overhead which will quickly consume the power of IoT devices. In addition, it will lead to transmission delays and low packet delivery ratio.

IV. Proposed Hopping Method

The smart channel hopping model (SCHM) is proposed for mitigating reactive jamming attacks in IoT networks. The formulation of the problem is modeled as the Markov Decision Process framework. The IoT devices interact with the environment and perform different actions in different states and eventually obtain best rewards. In contrast, the penalty reward will be allocated each time a bad action is taken. The next section summarizes the MDP framework for SCHM.

A. State Space

The legitimate IoT node can only observe the chosen channel for communication. Hence, the system's state space is illustrated using the following equation:

$$S = (c, j, q, e): c \in (0,1), j \in (0,1);\ q \in (0, \dots, Q);\ e \in (0, \dots, E)$$

where, c denotes the status of a radio channel, $c = 0$ signifies that a radio channel is idle and $c = 1$ signifies that the radio channel is busy. j denotes the status of the jammer, $j = 0$signifies that the jammer is not attacking a radio channel, $j = 1$ signifies that the jammer is attacking a radio channel. q denotes the number of data frames in the transmission queue, and e denotes the capacity of energy storage of the legitimate node. Q is the maximum data queue size, and E is the maximum capacity of energy storage. The system state space is then expressed as, $s = (c, j, q, e) \in S$.

B. Action Space

The legitimate node can execute any of the following actions, namely, remain idle, actively transmit frames, hop into another free radio channel, and transmit frames, when there is a jamming attack in the current selected channel: $A = (a: a \in \{1,2,3\}$.

a = { 1 the legitimate node remains idle,
2 the legitimate node transmits data,
3 the legitimate node hops into the new channel and transmits data.

C. Reward Function

The immediate reward for the system is defined in this section. It is described as the number of data frames that are sent to the receiver and positively acknowledged. The actions that the legitimate node takes must increase packet delivery ratio with minimum usage of energy and with less delays of the frames in the data queue.

D. Smart Channel Hopping Model

Double Deep-Q-Network (DDQN) is adopted for implementing SCHM. The DDQN reinforcement learning algorithm will be used to train the agent. The proposed model is shown in Fig. 3.

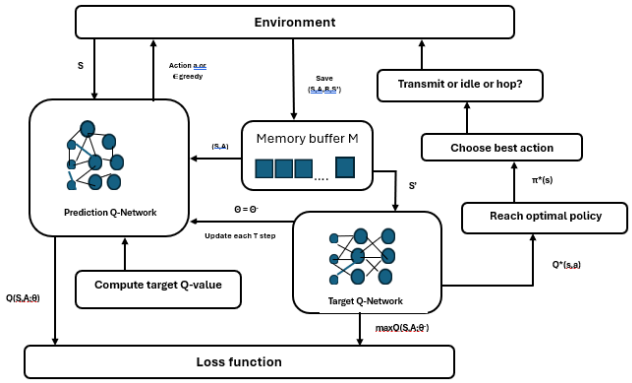


Fig. 3. Proposed smart channel hopping model.

The state space consists of four components, namely channel, jammer, buffer queue, and energy. The components form an important part of the actions that the agent will take during the mitigation process. The action space consists of three components, namely transmit, idle, and hop and transmit. The prediction Q-network, target Q-network, and the memory buffer are initialized. The agent observes the current state and takes the action, and then waits for the reward. Otherwise, it chooses the greedy action argmaxQ(s,a). The state-action pairs are stored in the memory buffer M for each episode. The mini-batch in the memory buffer is used to train the agent. The process continues until the network converges and an optimal policy is found.

V. CONCLUSION

In this research study, a systematic survey of literature review of methods for mitigating reactive jamming attacks in IoT networks was conducted. From the onset, the technologies that support IoT networks were presented. The conventional methods of defending against jamming attacks were discussed. That was followed by the discussion on how the popular online digital libraries were used to obtain conference papers and journal articles related to mitigating reactive jamming attacks. The research questions were raised regarding what the strengths and weaknesses of the existing mitigation methods are. In conclusion, this research study answered the questions.

This research study categorized the jamming mitigation methods into six, namely: Machine learning (ML)-based, deception-based, statistics-based, radio frequency-based, game theory-based, and encryption-based. The mitigation methods were further compared using the metrics, namely, computational overhead, defense response, and resource efficiency. The results in this research study suggest that most mitigation methods have strength of accuracy in detecting jamming attacks. Furthermore, the mitigation methods suffer from increased power consumption, less packet delivery ratio, and prolonged transmission delays.

The limitation of this SLR study is that the search strategy was manual and restricted to journal papers and conference papers. In the near future, the plan is to perform a broad automated search. Another limitation is the use of simplified search keywords for journal papers and conference papers. The use of search keywords can be generalized.

In retrospect, it is required to develop an efficient approach which detects and defends against reactive jamming attacks. As a future work, this research study proposed a smart channel hopping model for mitigating reactive jamming attacks in IoT networks. The model intends to reduce the power consumption of IoT devices, minimize packet errors, minimize packet loss, and minimize transmission delays, which will improve the packet delivery ratio. Furthermore, the research community will have a full knowledge of the limitations of existing RJAs mitigation techniques and will facilitate the advancement of improved mitigation techniques. Lastly, similar SLR studies can be considered for IoT cloud systems for addressing cybersecurity concerns and performance issues.

REFERENCES

[1] BUSINESS INSIDER (2020), Smart farming in 2020: How IoT sensors are creating a more efficient precision agriculture industry. Accessed: May 2025. [Online].Available at: https://www.businessinsider.com/smart-farmingiot-agriculture, 2020.

[2] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities". IEEE IoT journal, 2014.

[3] SEMTECH Smart healthcare. Accessed: May 2025. [Online]. Available at: https://www.semtech.com/lora/lora-applications/smarthealthcare, 2024.

[4] STATISTA (2021), Number of Internet of Things (IoT) connected devices. Accessed: May 2025. [Online]. Available at: https://www.statista.com/statistics/1183457/iotconnected-devices-worldwide/, 2021.

[5] A. Moussa, and I. Jabri, "Impact of RTS/CTS jamming attacks in IEEE 802.11ah dense networks". International Wireless Communications and Mobile Computing (IWCMC), 2021.

[6] M. Abdollahi, K. Malekinasab, W. Tu, and M. Bag-Mohammadi, "Physical-Layer Jammer Detection in Multihop IoT Networks". IEEE Internet Of Things Journal, Vol. 10, No. 23, 1 December 2023.

[7] A.S. Ali, S. Naser, and S. Muhaidat, "Defeating Proactive Jammers Using Deep Reinforcement Learning for Resource-Constrained IoT Networks". IEEE 34th Annual International Symposium on PIMRC, 2023.

[8] N. Huynh, D.T. Hoang, D.N. Nguyen, and E. Dutkiewicz, "DeepFake: Deep Dueling-Based Deception Strategy to Defeat Reactive Jammers". IEEE Transactions on Wireless Communications, VOL. 20, NO. 10, OCTOBER 2021.

[9] R.E. Navas, F. Cuppens, N.B. Cuppens, L. Toutain, and G.Z Papadopoulos, "Physical resilience to insider attacks in IoT networks: Independent cryptographically secure sequences for DSSS anti-jamming". Computer Networks Volume 187, 14 March 2021.

[10] N. Ahmed, D. De, F.A. Barbhuiya, and I. Hussain, "MAC Protocols for IEEE 802.11ah-Based Internet of Things: A Survey". IEEE INTERNET OF THINGS JOURNAL, VOL. 9, 2022.

[11] L. Arcangeloni, E. Testi, and A. Giorgetty, "Detection of Jamming Attacks via Source Separation and Causal Inference". IIE Transactions on Communications, Vol 71, No. 8, 2023.

[12] A.A. Fadele, M. Othman, I.A. Hashem, I. Yaqoob, M. Imran, and M. Shoaib, "A novel countermeasure technique for reactive jamming attack in internet of things", Multimedia Tools and Applications (2019) Vol 78, pp. 29899–29920, September 2019.

[13] A. Gouissem, K. Abualsaud, E. Yaacoub, T. Khattab, and M. Guizani, "IoT Anti-Jamming Strategy Using Game Theory and Neural Network". International Wireless Communications and Mobile Computing (IWCMC), 2020.

[14] D.T. Hoang, D.N. Nguyen, M.A. Alsheikh, S. Gong, E. Dutkiewicz, D. Niyato, and Z. Han, 2020. "Borrowing Arrows with Thatched Boats": The Art of Defeating Reactive Jammers in IoT Networks. Advances in Security And Privacy in Emerging Wireless Networks, 2020.

[15] A. Hussain, N. Abughanam, J. Qadir, and A. Mohamed, "Jamming Attacks in IoT Wireless Networks: An Edge-AI Based Approach", IoT

'22: Proceedings of the 12th International Conference on the Internet of Things pp. 57–64, 2022

[16] N. Huynh, D.T. Hoang, D.N. Nguyen, E. Dutkiewicz, and M. Mueck, "Defeating Smart and Reactive Jammers with Unlimited Power". IEEE on WCNC), DOI: 10.1109/ICC42927.2021.9500391 Jun 2021.

[17] G. Lee, and M. Kim, "Interference-aware Self-optimizing Wi-Fi for high Efficiency Internet of Things in dense Networks". Comput Commun., vols. 89–90, pp. 60–74, Sep. 2016

[18] S.J. Lee, Y.R. Lee, S.E. Jeon, and I.G. Lee, "Machine learning-based jamming attack classification and effective defense technique". Computers and Security Journal, ElSevier, 2023.

[19] Y. Liu, N. QI, Q. Shi, L. Jia, W. Wang, and Z. Huang, "Anti-Reactive-Jamming Wireless System: a Strategy of "Masking". 13th International Symposium on Antennas, Propagation and EM Theory (ISAPE), 2021.

[20] S. Nan, S. Brahma, C.A. Kamhoua, and N.O. Leslie, "Mitigation of Jamming Attacks via Deception". IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications: Track 2: Networking and MAC, 2020.

[21] X Tang, P. Ren, and Z. Han, "Jamming Mitigation via Hierarchical Security Game for IoT Communications". IEEE Access, Journal article, volume 6, 2018.

[22] N. Oukas, M. Boulif, and A. Abbas, "Mitigating Jamming Attacks in IoT RF-Devices through Dynamic Channel Hopping: A Novel Petri-nets Formulation". International Conference on Computer and Applications (ICCA), 2023.

[23] H. Pirayesh, and H. Zeng, "Jamming Attacks and Anti-Jamming Strategies in Wireless Networks: A Comprehensive Survey". IEEE Communications Surveys & Tutorials, Volume: 24, Issue: 2, Jan 2021.

[24] A. Pourranjbar, G. Kaddoum, A. Ferdowsi, and W. Saad, "Reinforcement Learning for Deceiving Reactive Jammers in Wireless Networks". IEEE Transactions on Communications, 2021.

[25] M. Reynvoet, O. Gheibi, F. Quin, and D. Weyns, "Detecting and Mitigating Jamming Attacks in IoT Networks Using Self-Adaptation". IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C), 2022.

[26] A.A. Sharah, H.A. Owida, and T.A. Edwan, "Aggressive Jamming Attack in IoT Networks". 4th IEEE MENACOMM, DOI: 10.1109, 2022.

[27] J. Singh, I. Woungang, S.K. Dhurandher, and K. Khalid, "A jamming attack detection technique for opportunistic networks". Science Direct Access, Internet of Things Journal, 2022.

[28] E. Testi, L. Arcangeloni, and A. Giorgetti, "Machine Learning-Based Jamming Detection and Classification in Wireless Networks", WiseML'23: Proceedings of the 2023 ACM Workshop on Wireless Security and Machine Learning, Pages 39–44, June 2023.

[29] K. Thiha, B. Soong, V. Vaiyapuri, and S. Nadarajan, "A new method of defeating reactive jamming: Hardware Design Approach". 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), 2019.

[30] B. Upadhyaya, S. Sun, and B. Sikdar, "Machine Learning-Based Jamming Detection in Wireless IoT Networks", IEEE VTS Asia Pacific Wireless Communications Symposium (APWCS) DOI: 10.1109/APWCS46464, 2019

[31] W. Yin, P. Hu, H. Zhou, G. Xing, and J. Wen, "Jamming attacks and defenses for fast association in IEEE 802.11ah networks". Science Direct Access, Computer Networks Volume 208, 2022.

[32] F. Zahra, Y.S. Bostanci, and M. Soyturk, "Comparative Analysis of Deep Learning Models for Detecting Jamming Attacks in Wi-Fi Network Data", 12th IFIP/IEEE International Conference Evaluation and Modelling in Wired and Wireless Networks (PEMWN), 2023.

[33] L. Zhang, T. Mao, C. Zhang, Z. Xiao, and X. Xia, "Reactive Jamming Detection Based on Hidden Markov Model". IEEE/CIC International Conference on Communications in China (ICCC), 2022.

[34] Z. Chkirbene, R. Hamila, and A. Erbad, "Secure Wireless Sensor Networks for Anti-Jamming Strategy Based on Game Theory". International Wireless Communication and Mobile Computing (IWCMC), 2023.

[35] A. Gouissem, K. Abualsaud, E. Yaacoub, T. Khattab, and M. Guizani, "Accelerated IoT Anti-Jamming: A Game Theoretic Power Allocation Strategy". IEEE Transactions On Wireless Communications, Vol. 21, No. 12, December 2022.

[36] S. Brauer, A. Zubow, S.Z.M. Roshandel, and S. Mashhadi-Soh, "On Practical Selective Jamming of Bluetooth Low Energy Advertising". IEEE Conference on Standards for Communications and Networking (CSCN), 2016.

[37] N. López-Vilos, C. Valencia-Cordero, C. Azurdia-Meza, S. Montejo-Sánchez, and S.B. Mafra, "Performance Analysis of the IEEE 802.15.4 Protocol for Smart Environments under Jamming Attacks". Sensors, Vol 21, No. 12, Jun 2021.

[38] L. Avanco, A.E. Guelfi, E. Pontes, A.A.A. Silva, S.T. Kofuji, and F. Zhou, "An Effective Intrusion Detection Approach for Jamming Attacks on RFID Systems". International EURASIP Workshop on RFID Technology (EURFID), 2015.

[39] K Grover, A Lim, and Q Yang, "Jamming and anti-jamming techniques in wireless networks: A survey," Int. J. Ad Hoc Ubiquitous Comput., vol. 17, no. 4, pp. 197–215, 2014.

[40] S Vadlamani, B Eksioglu, H Medal, and A Nandi, "Jamming attacks on wireless networks: A taxonomic survey," Int. J. Prod. Econ., vol. 172, pp. 76–94, Feb. 2016.

[41] R Di Pietro and G Oligeri, "Jamming mitigation in cognitive radio networks," IEEE Netw., vol. 27, no. 3, pp. 10–15, May/Jun. 2013.

Virtual Assistant Based on Recurrent Neural Networks: Scope and Coverage of Health Insurance

Diego Alberto Paz-Medina[1], David Eduardo Rojas-Cavassa[2], Ernesto Adolfo Carrera-Salas[3]
School of Engineering, Universidad Peruana de Ciencias Aplicadas (UPC), Lima, Peru [1, 2, 3]

***Abstract*—In Peru, the use of health services provided by state institutions has decreased, largely due to perceived deficiencies in care quality, such as delays in medical attention and administrative barriers that hinder timely access to information on insurance scope and coverage. This study develops a web-based application with an integrated chatbot that leverages Artificial Intelligence (AI) and a Recurrent Neural Network (RNN) to provide accurate and accessible information about institutional insurance, including the characteristics of different insurance types and specific coverage cases. The application integrates a chatbot powered by an AI model based on Natural Language Processing (NLP) from OpenAI, with intent recognition handled by a dedicated classifier. In comparison to existing chatbots in the healthcare field, the model proposes a hybrid approach based on RNN and GPT-3.5, optimized for health insurance queries in public sector contexts. The system was evaluated with 50 representative questions scored on four criteria: clarity, relevance, coherence, and accuracy. The chatbot achieved an overall mean response accuracy of 82% and a mean user satisfaction score of 4.59/5, indicating strong acceptance and usability. These results suggest that the combination of these technologies constitutes an effective alternative for addressing queries relevant to users of state health systems.**

***Keywords*—Insurance; insurance coverage; web application; chatbot; Recurrent Neural Networks**

I. Introduction

According to the nominal registry of insured individuals published in early 2022, reports that 98.96% of Peru's total population has health insurance coverage [1], with the majority covered by the Comprehensive Health Insurance (SIS) and Peru's Social Health Insurance (EsSalud), accounting for 69.5% and 29.3%, respectively. As stated in [2], in 2022, SIS showed a steady increase in enrollees from 2009 to 2022, while EsSalud exhibited a relatively constant affiliation trend. However, between 2020 and 2022, there was a decrease in the EsSalud-insured population, dropping from 30.4% to 22.8% of the total Peruvian population. In [3], the authors highlighted that, due to poor management and weak governance in recent years, the institution has experienced instability, resulting in delays in project implementation, decision-making processes, and turnover in key administrative areas. Additionally, [3] notes that the percentage of EsSalud-insured individuals visiting hospitals declined from 45% in 2004 to 31% in 2022. This decline is largely attributed to a perceived reduction in care quality [4], including delays in medical and consultation services, as well as the imposition of time constraints or administrative barriers that hinder effective understanding and access to consultation services.

Among the various approaches explored in previous research, [5] presents the comparison of a chatbot based on a Sequential Matching Framework (SMF), supported by a Recurrent Neural Network (RNN) algorithm, with other chatbots based on Natural Language Processing (NLP), such as RootyAI, ChatterBot, and DeepQA. In study [6], an artificial intelligence model was developed, capable of understanding terms related to the insurance domain for the resolution of queries. Although [7] and [8] document the development of a chatbot, [7] features the use of different types of algorithms, which are more advanced, such as the use of DeepQA for the chatbot's construction and DeepProbe for rewriting queries. In study [8], a chatbot was developed locally in hospitals, supported by Artificial Intelligence (AI), for the resolution of doubts related to immunization held by the medical personnel in charge of minors. In [9], the authors present an algorithm for adapting GPT models to specific tasks by combining information from different sources. This improves the accuracy of the model's responses and makes it more useful for specific tasks. This differs from the proposed approach, which focuses on the field of health insurance for the general public, where the user is not required to have prior knowledge of specific related terms. Finally, [10] describes the development of a chatbot in Ethiopia for mental health support, using a hybrid architecture based on Seq2Seq models, Amharic embeddings, and deep neural networks (LSTM and GRU).

In the present study, a web-based application was developed incorporating an integrated chatbot designed to provide both EsSalud-insured and uninsured individuals with access to information regarding insurance modalities, specific cases in which insurance coverage may apply, and healthcare centers within this public institution that offer treatment for the conditions specified by users. Based on the queries submitted by the user, the system generates contextually appropriate responses using an Artificial Intelligence (AI) model previously trained by a Recurrent Neural Network (RNN) algorithm. The model was trained in data related to the scope and coverage of the various insurance plans offered by EsSalud. Through the implementation of this application, users, regardless of insurance status, are provided with a digital channel through which frequently asked questions about EsSalud insurance can be addressed, thereby reducing waiting times for in-person assistance and alleviating congestion at insured service centers.

During the validation phase, the system's response accuracy was first assessed at the backend level. Specific validation criteria were established, and the system's responses were evaluated against official information obtained from EsSalud. The evaluation indicated an overall mean response accuracy of

82%, with the main limitation being the model's difficulty in generating clear and concise responses. Usability testing was then conducted with 32 participants who interacted with the chatbot, followed by a 10-item questionnaire designed to evaluate user experience. The results showed a general satisfaction score of 4.59 out of 5 on a Likert scale, a dropout rate of 16%, and a perceived overall improvement of 78% compared with human agents (41% rated the system as "much better" and 37% as "better"). In conclusion, the findings support the efficacy of digital solutions, such as AI-driven chatbots, in efficiently disseminating basic insurance-related information and responding to common inquiries, thereby improving service accessibility and responsiveness for both insured and uninsured populations.

The structure of the study is organized as follows: Section II provides a review of related works; Section III introduces the concepts applied throughout the development process; Section IV describes the methodology employed; Section V details the results of the experiment conducted; Section VI discusses the findings; and Section VII presents the conclusions.

II. Related Works

In [5], the authors present a study on the response-generation strategies employed by existing chatbots, aiming to identify shortcomings in maintaining coherent dialogues with users. Subsequently, a domain-specific chatbot named IntelliBot is proposed, which is based on a Sequential Matching Framework (SMF). This framework transforms each word into a vector according to its contextual meaning and employs hidden states from Recurrent Neural Networks (RNNs) to determine relationships between expressions. Additionally, the study examines alternative approaches for constructing conversational systems, including language generation techniques and neural network-based retrieval models. The model was trained and evaluated using three experiments and datasets. The first experiment utilized a movie dialogue dataset with over 200,000 samples and a batch size of 250 records. The second experiment used a credit card insurance dataset comprising over 10,000 samples with a batch size of 150 records. The third experiment involved a dataset of insurance-related terms and keywords with 1,000 samples and a batch size of 50 records. Following the execution of 71 questions across seven different categories, the results demonstrated the superiority of IntelliBot in terms of accuracy and memory, achieving a score of 0.97. In contrast, traditional chatbots obtained lower performance scores: 0.66 for RootyAI, 0.71 for ChatterBot, and 0.93 for DeepQA. However, despite these promising results, IntelliBot focuses primarily on general insurance contexts, without addressing the accessibility barriers or usability needs specific to public health institutions. Unlike the previous approach, the proposed model integrates a hybrid architecture optimized for state health insurance queries, user satisfaction, and accessibility for non-specialized users. This highlights a research gap regarding the adaptation of AI-driven conversational systems to public-sector health environments.

In study [6], an artificial intelligence model was developed using a Recurrent Neural Network (RNN) and a Convolutional Neural Network (CNN) to improve the resolution of queries related to the insurance domain. This model was created in response to the growing demand from individuals seeking information about the types of insurance they may obtain. Three datasets were used for training, validation, and testing, comprising a total of 16,889 questions, 16,889 correct answers, and 168,890 incorrect answers. The developed model demonstrated a higher accuracy rate than the baseline model, achieving a precision of 78.34%, which was attributed to its enhanced semantic understanding capabilities. This study addresses the same thematic area as the project under development, although it focuses on all types of insurance, whereas our work concentrates specifically on health insurance. Additionally, a key distinction lies in the response mechanism provided to users: while the referenced study focuses solely on the development of the model, the proposed project involves the implementation of the model within a digital alternative-specifically, a chatbot integrated into a web-based platform.

In [7], the authors present a design and implementation model for an intelligent chatbot using advanced algorithms. To achieve this, a high-quality dataset was required to train deep learning models effectively. The proposed model employed Long Short-Term Memory (LSTM) neural networks and the Sequence-to-Sequence (Seq2Seq) algorithm, along with the Bag of Words (BoW) model, beam search decoding for response prediction, and the BLEU-N algorithm to evaluate the chatbot's accuracy. As a result, the proposed method was compared and analyzed using the BLEU-N score against a traditional chatbot, a DeepQA-based chatbot, and rewritten versions of DeepProbe. The findings showed that the proposed method outperformed the others with a 59.44% improvement in BLEU-N score and overall performance. However, this approach focuses mainly on linguistic precision, without addressing user experience or applicability in public health contexts. The proposed model builds upon these advances by adapting deep learning methods to domain-specific insurance queries, helping to close this gap.

In study [8], an AI-assisted chatbot was developed to address questions related to childhood immunization in a low-resource and low-literacy setting in Karachi, Pakistan. The chatbot employed a continuous learning process through a Machine Learning algorithm integrated with a "Human-in-the-Loop" function, which aims to replicate natural human conversations without relying solely on keyword matching. Among the total number of inquiries (874), the most frequently addressed topics were vaccine expiration (32.4%), delays in vaccination due to child illnesses (16.8%), and management of side effects (15.7%). Additionally, it was found that most questions were answered promptly and in a manner that was clear to users. Regarding user acceptance, 90% indicated they would have no issue continuing to use the system; however, 40% expressed they would feel more comfortable knowing the source of the responses. Compared to the chatbot described in that study, the present research aims to address users' questions even when they have limited knowledge of the specific domain-namely, the scope and coverage of health insurance. This helps bridge the knowledge gap necessary for using a digital tool, as the user can make inquiries either in formal or colloquial language, both of which are interpreted in the same way thanks to the selected AI model.

In [9], the authors propose a method for adapting GPT-based artificial intelligence models to specific scenarios, such as patent claim generation, multilingual text summarization, and answer generation in question-answering systems. The new model, named Dempster-Shafer Theory (DST) framework, performs information fusion based on Dempster's theory of evidence, combining inputs from multiple sources to enhance accuracy and decision-making in response generation. The results showed that the proposed method improved the contextual accuracy of responses by 8.95% and 7.56%, respectively, compared to those generated by GPT-2 and GPT-3.5 models. However, its application remains focused on technical and text-processing domains, without exploring user-centered implementations in social or public service contexts. The present study addresses this limitation by applying GPT-based reasoning to healthcare insurance queries, emphasizing usability and accessibility.

In [10], the authors propose the design and implementation of an artificial intelligence-based chatbot called "PsychoBot", aimed at assisting users with potential mental health disorders in Ethiopia. For natural language processing, Word2Vec techniques (skip-gram and CBOW) were employed, along with Recurrent Neural Networks, including variants such as LSTM, GRU, double GRU, and transposed GRU. The architecture identified as optimal was the transposed GRU, due to its superior performance in terms of accuracy (79.86%), lower validation loss (0.208), and reduced training time (1434 minutes). In contrast to the present study, which focuses on the domain of health insurance, this work concentrates on psychological diagnostics, emphasizing the implementation in the local language and its robustness in handling diverse user queries. However, it does not integrate a web-accessible platform or a user-friendly chatbot interface, features that are essential to improving the usability of the developed solution and enhancing the overall user experience.

Table I presents a comparison of the methods used, datasets, and evaluation metrics across the related works.

TABLE I. COMPARISON OF RELATED WORKS AND THE PROPOSED MODEL

Year	References	Description of method/technique	Dataset size	Performance metric/s
2020	Paper 1 [5]	The working styles of chatbots are examined through a model based on Recurrent Neural Networks (RNN) and the Semantic Matching Framework (SMF).	71 questions across seven different categories.	Kappa coefficient and F1 score.
2023	Paper 2 [6]	The MFBT model, which focuses on insurance types, is developed using artificial intelligence with a BiLSTM-TextCNN architecture.	A total of 16,889 questions, divided into training, validation, and testing sets.	Model precision, recall, and F1 score.
2023	Paper 3 [7]	A design and implementation for an intelligent chatbot is presented, combining LSTM, a Seq2Seq architecture with attention mechanism, data preprocessing, Bag of Words (BoW) model, and supervised training.	Dialog dataset, containing 3,725 question–answer pairs.	BLEU score (Bilingual Evaluation Understudy).
2024	Paper 4 [8]	Bablibot: a chatbot designed to answer questions related to childhood immunization, employing a continuous learning process through Machine Learning with a "Human-in-the-Loop" function.	In Karachi (Pakistan), a thematic analysis of interviews was conducted with 2,202 caregivers, of whom 677 (30.7%) interacted with Bablibot.	User acceptance based on interview feedback.
2024	Paper 5 [9]	A method for adapting GPT models to various real-world scenarios to enhance performance in contextual dialogue classification.	Uses the Dialog dataset (3,725 Q&A pairs), DailyDialog (13,118 dialogues from everyday interactions), and WikiQA (3,047 questions sourced from Bing query logs).	BLEU score and cosine similarity score.
2025	Paper 6 [10]	A mental health support chatbot based on RNNs (LSTM, GRU), Seq2Seq architecture, Amharic word embeddings (Skip-gram and CBOW), and memory networks, combining generative and retrieval-based approaches in a hybrid architecture.	Employs a dataset compiled from the DSM-5 manual, client case histories, online psychological counseling websites, and general everyday conversations.	Precision, recall, and F1 score.
2025	The model proposed	A chatbot was developed integrating communication between an AI model and an RNN to address inquiries about the scope and coverage of EsSalud health insurance.	109 questions related to insurance services provided by EsSalud only for training purposes.	Average response accuracy rate and user survey results.

Source: Own Elaboration.

III. CONTRIBUTION

This section defines and contextualizes the key concepts underlying the proposed research. Each concept provides a basic definition and serves as a fundamental element for the design and implementation of the model.

A. Health Insurance

A contract established with an insurance company through the payment of a monthly premium. In exchange for the premium, the insurance company agrees to cover a variety of medical services, depending on the applicant's plan [11]. This concept was essential for structuring the chatbot's knowledge base. Based on the institutional documents collected, the system was trained to identify and respond to frequently asked questions regarding health insurance. A Recurrent Neural Network (RNN) was employed to accurately classify user queries, while the AI language model was responsible for generating clear and comprehensible responses. In contrast to previous chatbots that focused mainly on general healthcare information, the proposed model offers a domain-specific approach to health insurance, capable of delivering policy-based and context-aware responses aligned with institutional regulations.

B. Insurance Coverage

It refers to the protection provided by a health insurance policy against unexpected events, preventing these from becoming a significant financial burden in the present or future [12]. In the system's development, this concept was essential

for structuring the chatbot's responses regarding the services guaranteed by each type of insurance. Institutional documents in PDF format, which detail the coverage information, were used to fine-tune RNN algorithm for pattern recognition in user queries. Meanwhile, the OpenAI language model complements these responses by expanding the contextual information, helping users better understand which services are included in their health insurance plan. Unlike conventional systems with static information, the proposed model retrieves and integrates content directly from institutional documents to generate dynamic, personalized, and policy-specific responses, ensuring greater accuracy and relevance.

C. Insurance Scope

The limits of protection offered by an insurance policy in the event of unforeseen circumstances [13]. The concept of scope was important for training the chatbot to explain the limits and conditions of each insurance plan. This information, extracted from official documents, is processed by the RNN to accurately categorize user queries, while the AI model generates responses that clarify the extent of insurance coverage in various situations. Unlike traditional models based on predefined answers, the proposed system interprets each policy's scope dynamically, offering clearer, adaptive, and user-tailored explanations.

D. Artificial Intelligence

A scientific field focused on the development of computers and machines capable of thinking, learning, and making decisions in a manner similar to humans, or of processing data sets that exceed human analytical capacity [14]. The system is implemented using OpenAI's GPT-3.5 Turbo model, which interprets complex queries related to health insurance and generates coherent responses in natural language. This enables the chatbot to provide fast and accessible assistance without the need for human intervention. Unlike previous systems that used AI mainly for classification or retrieval tasks, this model employs generative AI to produce contextually accurate and user-friendly responses, improving accessibility for non-technical users.

E. Recurrent Neural Networks

A machine learning model based on neural networks that processes sequential data such as text or time series, converting it into specific sequence-based outputs [15]. They are used to process sequential health insurance queries, categorizing them based on user intent. A simple RNN architecture was employed due to its efficiency and speed in handling short sequence tasks, such as brief user queries. The aim is to provide a perspective on how this type of algorithm can "communicate" with others to enhance the responses it generates.

F. Web Application

A software program that runs directly on a web browser via the internet [16]. In the present work, this concept was applied when defining the appropriate medium for developing a solution that allows users to make inquiries about health insurance and access related information without requiring any external installation, by simply using their preferred web browser. Compared to the previously mentioned chatbots, the proposed solution aims to enable communication with the chatbot through an easily accessible web page, featuring a responsive design tailored to the user's needs.

IV. Method

This section presents the developed method according to the stages outlined in Fig. 1:

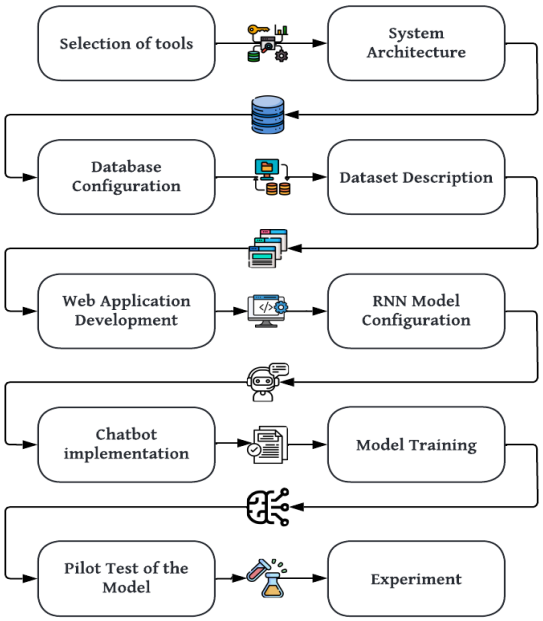


Fig. 1. Stages of the proposed method. Source: Own elaboration.

A. Selection of Tools

Regarding the AI provider, OpenAI was selected based on API maturity, tooling, and empirical evaluations of model quality reported in recent benchmarks [17]. GPT-3.5 Turbo was used for generating the chatbot's responses. Additionally, OpenAI was chosen for its widespread popularity in projects requiring human-like responses and for its ease of integration through the use of an API key, which grants access to OpenAI's models. Amazon Web Services (AWS) was chosen as the cloud service provider due to the broad range of services it offers for deploying web applications, machine learning resources, and its cost-effectiveness, making it the most suitable option considering the project's needs and constraints. Python was chosen as the programming language for data processing due to its robust ecosystem for NLP and machine learning, its ease of use, and its efficiency in handling large-scale datasets.

During the selection process for the most suitable AI model provider, four options were compared: OpenAI, Google Cloud AI, IBM, and Microsoft. The evaluation was based on five key criteria identified as most relevant in [18]: model governance, performance, ease of integration, scalability, and cost. OpenAI stood out for its high performance and scalability, being highly efficient in natural language processing tasks and capable of handling large data loads. Google Cloud AI provides robust tools but requires additional configurations to reach optimal performance. IBM facilitates model integration but tends to be more costly. Microsoft offers a wide range of options, though its implementation may require advanced knowledge and complex configuration.

Based on a qualitative analysis, the authors assigned scores to each provider according to the evaluation criteria. These scores were then weighted using a prioritization matrix that reflects the relative impact of each criterion. The results are presented in Table II, with OpenAI achieving the highest score of 3.94 as the most suitable AI model provider.

Following an evaluation of different database management systems, MySQL was selected due to its open-source nature. Unlike SQL Server, MySQL allows free usage and broad customization, as noted in [19]. It is recognized for its ease of installation and use, as well as its compatibility with multiple programming languages and operating systems. Moreover, it delivers optimal performance in data handling, enabling fast queries and updates. This makes it suitable for the current project, which does not require intensive performance, especially considering its lower resource requirements and ease of use without complex configurations. Additionally, its compatibility with Railway allows for seamless cloud deployment, accelerating testing and project management.

TABLE II. COMPARISON OF ARTIFICIAL INTELLIGENCE MODEL PROVIDERS

Criteria	*OpenAI*	*Google Cloud AI*	*IBM*	*Microsoft*
Model Governance	0.75	0.56	0.75	0.56
Performance	1.00	0.75	0.75	1.00
Ease of Integration	0.38	0.63	0.50	0.25
Scalability	1.25	0.75	1.00	0.75
Cost	0.56	0.56	0.56	0.56
Total	3.94	3.25	3.56	3.13

Source: Own Elaboration.

For this study, a simple Recurrent Neural Network (RNN) architecture was selected, thanks to its lightweight structure, quick training process, and suitability for short sequences-such as common questions related to health insurance. Unlike LSTM and GRU architectures, which are designed to capture long-term dependencies and process extended sequences, the simple RNN proved sufficient for our use case, where the inputs are brief and do not require long-term memory. While LSTM and GRU architecture are optimal for complex tasks, they involve a larger number of parameters, longer training times, and greater computational resource usage, as noted in [20]. Likewise, although Transformer models excel in complex tasks and large datasets, they involve high computational costs and require substantial data and specialized hardware, making them unfeasible for this study [21]. In contrast, the simple RNN offers a lighter and easier-to-implement design, which is essential to maintaining model simplicity without compromising its response generation capabilities. This comparison is shown in Table III, where an evaluation using a Likert scale from 1 to 5 was conducted. The Simple RNN achieved the highest overall score of 3.67, thus selected as the RNN architecture used in the proposed model.

TABLE III. COMPARISON OF RECURRENT NEURAL NETWORK ARCHITECTURES

Criteria	*Simple RNN*	*LSTM*	*GRU*	*Transformer*
Training Time	5	2	3	4
Long-term dependency capture	5	2	3	2
Number of parameters / complexities	1	5	4	5
Performance on short sequences	4	4	4	4
Training stability	2	5	4	4
Hardware requirements	5	2	3	2
Total	**3.67**	3.33	3.5	3.5

Source: Own Elaboration.

B. System Architecture

Fig. 2 presents the integrated system architecture, which follows a multilayer structure composed of the Presentation Layer, Application Layer, Data Layer, and Intelligent Service Layer; ensuring modularity, scalability and secure communication between all components.

Presentation Layer: Developed in Angular and hosted on AWS Amplify, this layer allows users to interact with the chatbot through a responsive web interface accessible from desktop and mobile devices. User queries are transmitted securely to the backend via RESTful APIs.

Application Layer: Implemented in Python (Django) and deployed on Railway Cloud, this layer processes user inputs, manages communication between the front end, the RNN algorithm, and the AI model. It also handles intent classification and message routing to ensure relevant and coherent responses.

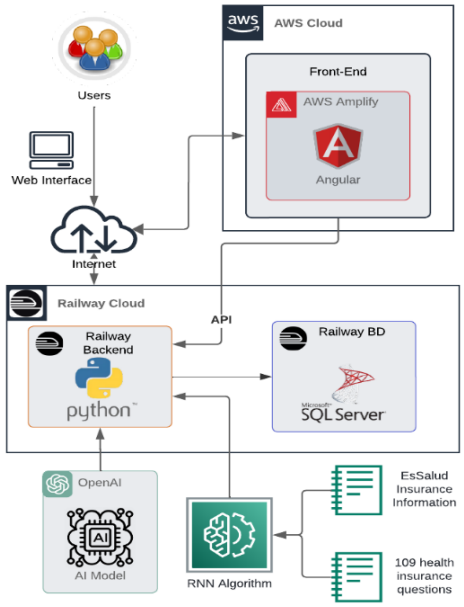


Fig. 2. System architecture of the proposed method.
Source: Own elaboration.

Data Layer: Hosted on Railway BD using Microsoft SQL Server, this layer stores insurance information, user interactions, and the dataset of 109 representative questions used for RNN training, enabling dynamic query retrieval and model improvement.

Intelligent Service Layer: Integrates the external OpenAI model with the internal RNN algorithm. The RNN identifies user intent, and OpenAI generates the final natural language response, combining deterministic classification with generative reasoning.

Interaction Flow: Users submit queries via the web interface; the backend classifies them using the RNN, retrieves related information from the database, and requests a response from the OpenAI model. The generated answer is then returned to the user interface, completing the interaction cycle efficiently.

C. Database Configuration

Since Railway was selected as the service provider, which allows integration with different database engines such as MySQL Server, this was used for storing the data related to the application. In addition, the same MySQL Server application was used to manage the data, both for queries and for editing. This, in turn, connects with the web application services to perform the queries required by the system.

A database model was developed using a class diagram and an entity-relationship model, having mainly the entities User, where personal data such as full name and login credentials will be stored, Responses, where the responses and the rating given by the user are stored to identify whether the response is accepted or rejected by the user in cases where more specialized attention is required, and Insurance, data related to the person's insurance, as well as the validity of the coverage.

D. Dataset Description

For the training and validation of the Recurrent Neural Network (RNN) model, a dataset was constructed from multiple sources containing information related to health insurance plans. Five PDF documents were used, each corresponding to a different type of health insurance. These documents included detailed information on the characteristics, coverages, exclusions, and validity periods of each insurance plan.

Additionally, a complementary PDF document containing 109 pairs of frequently asked questions (FAQs) and their corresponding answers was utilized. These questions were derived from real user inquiries and were validated and reviewed by personnel from the insurance department to ensure their accuracy and consistency. The questions were specifically designed to train the model in natural language understanding. To mitigate the limitations of the dataset, data augmentation techniques were applied, such as synonym replacement, paraphrasing, and back-translation, to increase linguistic variability while preserving semantic meaning.

Prior to model training, the question–answer pairs underwent preprocessing steps, including the removal of redundant symbols, text normalization (conversion to lowercase), and structuring into a suitable format for model input. The data were then tokenized, encoded, and used to train the RNN model. Through this process, the model learned to associate specific types of user queries with their appropriate responses based on the patterns within the dataset.

The outcome of this training provides a solid foundation for the chatbot to interpret user questions and generate coherent responses consistent with official information regarding the scope and coverage of health insurance plans.

E. Web Application Development

With respect to the visual layer for the user, it was decided to develop it using the Angular framework, given the development team's experience in programming business logic and modeling modules in TypeScript. This layer will be hosted on the AWS Amplify service due to its support for web applications, its ease of use for deployment, and its integration capabilities with other data processing services provided by AWS [22].

Its deployment will be carried out using Railway services, which will provide the functions performed in the backend through an API Gateway, which communicates with both the presentation layer on AWS Amplify and the artificial intelligence model provided by OpenAI [23]. To facilitate portability, scalability, and execution environment management, the backend has been developed and deployed within Docker containers, allowing for more efficient dependency control and greater consistency between development (localhost) and production (cloud) environments.

Communication begins with the front end in Amplify, sending data in JavaScript Object Notation (JSON) format to the backend, which is hosted on Railway services and receives the information through the API Gateway, then forwards this information to the OpenAI AI model [24].

F. RNN Model Configuration

The model architecture is based on a simple Recurrent Neural Network (RNN). Development and training were conducted in a local environment equipped with an Intel Core i5-6600 processor, 16 GB of RAM, and an AMD RX-480 dedicated graphics card. The system operated on Windows 10 Pro, using Visual Studio Code as the development environment, TensorFlow 2.15 as the deep learning framework, and Python 3.10 as the programming language.

During the experimental tests, equivalent SimpleRNN, LSTM and GRU models were trained under the same configuration parameters (number of neurons, embedding size, and sequence length). Although their accuracy was similar, the SimpleRNN achieved comparable performance with lower training time and memory usage, representing a more efficient cost–benefit balance for this system. Furthermore, since the input data do not involve long-term temporal dependencies, the use of LSTM or GRU would not significantly improve response quality, making the SimpleRNN a sufficient and lightweight solution.

The implemented architecture consists of three main layers: an Embedding layer, which transforms words into 64-dimensional numerical vectors; a SimpleRNN layer with 128 units, responsible for learning the sequential relationships

between words; and a dense output layer with a softmax activation function, which predicts the most probable response among the possible output classes.

Training was performed using the Adam optimizer with a learning rate of 0.001, employing sparse categorical crossentropy as the loss function and accuracy as the performance evaluation metric. The dataset was divided into 80% for training and 20% for testing, with a maximum sequence length of 20 tokens and a vocabulary size limited to 10,000 words.

The model was trained for 10,000 epochs, achieving stable convergence, and an adequate level of accuracy on the training data. Finally, the trained model, along with the tokenizer and label encoder, were stored in reusable formats, enabling their subsequent loading and integration into the web application.

The RNN model processes the user query through tokenization, padding, and intent classification, then integrates the result with the language generator. The full operational sequence is illustrated in Appendix A, which presents the pseudocode of the proposed model.

G. Chatbot Implementation

The application's front end, developed in AWS Amplify, was integrated with the backend services developed in Railway. Also, tests were conducted related to prompt construction to evaluate the request and response flow with the AI model. The backend services are hosted on Railway as the main server. For this purpose, Docker containers were used: one dedicated to the OpenAI-based chatbot, and another for the Recurrent Neural Network (RNN) model. Tests were conducted related to prompt construction to assess the request handling and the quality of the responses generated by the AI and RNN models.

Fig. 3. Chatbot view of the web application. Source: Own elaboration.

The main view of the developed chatbot is shown in Fig. 3. Users can submit queries regardless of whether they are registered, by entering a message in the input bar at the bottom, within the maximum character limit. Additionally, two options are provided for interacting with the chatbot. The first, "What type of insurance do I need?", displays a series of predefined questions and answers, which lead to a recommendation for one of the five available insurance plans offered by the institution, based on the user's selection. The second option, "Frequently Asked Questions", presents a list of stored questions and provides an answer upon selection, which is then sent to the endpoint for communication with the chatbot and the RNN algorithm.

H. Model Training

During the training phase, documentation related to the five types of insurance provided by EsSalud was collected, covering information about their coverage and the requirements for enrollment. These data were used to train both the RNN model and the external AI model provided by OpenAI. The main chatbot was built using language models from the same provider, responsible for generating dynamic and understandable responses to freely formulated user queries, with its responses strictly to the different types of health insurance. On the other hand, the RNNs were previously trained with specific data related to the insurance types. The objective of this model is to analyze and classify preconfigured or frequently asked questions (109 questions and answers), which require technical responses based on learned rules or patterns. The RNN serves as a specialized support system that enables fast and accurate responses without requiring as many resources as the main chatbot.

I. Pilot Test of the Model

Subsequently, a pilot test was conducted at the backend of the developed application, specifically focusing on the API used to communicate with both the RNN and AI models. For the test, a list of questions was prepared along with the corresponding answers obtained from EsSalud's insured services personnel. A total of 50 questions, along with their respective answers (distinct from the 109 used to train the model), were formulated and then compared with those generated by the developed model, using four criteria derived from those in [25]: Clarity (whether the response focuses on what is being asked without unnecessary detours), Relevance (whether the response is relevant to the context of the question), Coherence (whether the information is consistent with the question and logically coherent), and Accuracy (whether the response matches the verified information). Each criterion was assigned a weight of 25%, aiming to achieve an overall compliance rate above 80%. The comparison process was conducted by the development team, with support from personnel in the insurance systems department of the selected institution.

Table IV presents the results of the 50 questions posed to the AI model provided by OpenAI alongside the RNN model, evaluating the responses according to the established criteria.

In the responses obtained for the insurance-related questions, it was found that the answers were often not very clear, as they tended to include more information than necessary. For this reason, these questions had a compliance percentage ranging from 50% to 75%.

TABLE IV. CALCULATION OF THE RESPONSE ACCURACY PERCENTAGE OF THE AI MODEL

Questions	*Criteria*				*%*	*Questions*	*Criteria*				*%*
	Clarity	Relevance	Coherence	Accuracy			Clarity	Relevance	Coherence	Accuracy	
1	Yes	No	Yes	No	50%	**26**	Yes	No	Yes	No	50%
2	Yes	Yes	Yes	Yes	100%	**27**	Yes	Yes	Yes	Yes	100%
3	Yes	Yes	Yes	Yes	100%	**28**	Yes	Yes	Yes	Yes	100%
4	No	Yes	Yes	No	50%	**29**	No	Yes	Yes	Yes	75%
5	Yes	Yes	Yes	Yes	100%	**30**	Yes	Yes	Yes	Yes	100%
6	Yes	Yes	Yes	No	75%	**31**	No	Yes	Yes	No	50%
7	Yes	Yes	Yes	Yes	100%	**32**	No	Yes	Yes	No	50%
8	Yes	Yes	Yes	Yes	100%	**33**	Yes	Yes	Yes	Yes	100%
9	Yes	Yes	Yes	Yes	100%	**34**	No	Yes	Yes	Yes	75%
10	Yes	No	Yes	Yes	75%	**35**	No	Yes	Yes	No	50%
11	Yes	Yes	Yes	Yes	100%	**36**	Yes	Yes	Yes	Yes	100%
12	Yes	No	Yes	Yes	75%	**37**	Yes	Yes	No	No	50%
13	Yes	No	Yes	No	50%	**38**	Yes	Yes	Yes	No	75%
14	Yes	No	Yes	No	50%	**39**	Yes	Yes	Yes	No	75%
15	Yes	Yes	Yes	Yes	100%	**40**	Yes	Yes	Yes	Yes	100%
16	Yes	Yes	Yes	Yes	100%	**41**	Yes	No	Yes	No	50%
17	No	Yes	Yes	Yes	75%	**42**	Yes	Yes	Yes	No	75%
18	No	Yes	Yes	Yes	75%	**43**	Yes	Yes	Yes	No	75%
19	Yes	No	Yes	Yes	75%	**44**	Yes	Yes	Yes	Yes	100%
20	Yes	Yes	Yes	Yes	100%	**45**	Yes	Yes	Yes	No	75%
21	Yes	Yes	Yes	No	75%	**46**	Yes	Yes	Yes	Yes	100%
22	Yes	Yes	Yes	Yes	100%	**47**	Yes	Yes	Yes	Yes	100%
23	Yes	Yes	Yes	Yes	100%	**48**	Yes	Yes	Yes	No	75%
24	Yes	Yes	Yes	Yes	100%	**49**	Yes	Yes	Yes	Yes	100%
25	Yes	Yes	No	Yes	75%	**50**	Yes	Yes	Yes	Yes	100%

Source: Own Elaboration.

TABLE V. RESULTS OF THE AI MODEL EVALUATION

Value	
Mean	82%
Median	75%
Mode	100%
Standard Deviation	0.196
Minimum	50%
Maximum	100%

Source: Own Elaboration.

Table V presents the statistical data related to the responses provided by the artificial intelligence model alongside the RNN model. The mean accuracy was 82%, indicating a high level of acceptance, as it falls above the 70% threshold. The data show moderate dispersion around the mean, meaning the values are not highly spread out, but there is still noticeable variability. A mode of 100% was observed, making it the most frequent accuracy value among the responses. The answers fall within the upper range, with a minimum of 50% and a maximum of 100%, indicating that all responses had an accurate percentage above 50%.

A 95% confidence interval was calculated for the model's accuracy using the standard normal approximation, resulting in 0.82 ± 0.0543 (0.7657–0.8743). This interval confirms, with 95% confidence, the reliability and stability of the model's performance across the evaluated dataset.

J. Experiment

A test of the application was conducted with a group of 32 individuals who were either insured by EsSalud or interested in obtaining insurance from the institution. The users interacted

with the chatbot, powered by an RNN algorithm and an AI model, to make various insurance-related inquiries.

The participants were selected through a non-probabilistic sampling approach. The group consisted of 12 women and 20 men, all adults and active users of the national health insurance system. Participants were chosen based on their availability and familiarity with digital platforms, allowing an adequate representation of real insured users. The sample composition reflects the typical demographic distribution of the target population. This approach helps minimize potential bias and provides valid insights into user interaction with the proposed system.

Participants were then asked to complete a ten-question survey regarding their user experience, assessing their satisfaction with certain features or any difficulties encountered during usage. Table VI presents the scale and metric associated with each question.

TABLE VI. QUESTIONS ASKED IN THE USER SURVEY

No.	*Question*	*Scale*	*Metric*
Q1	How accurate do you consider the responses from the chatbot?	Likert	Average perceived accuracy by users
Q2	How often were the chatbot's responses useful and relevant to your inquiries?	Likert	Frequency of relevant vs. irrelevant responses
Q3	Are you satisfied with the time it takes for the chatbot to respond?	Likert	Satisfaction with response time and average response times
Q4	How easy was it to interact with the chatbot?	Likert	Average satisfaction with ease of use
Q5	What is your overall satisfaction level with the chatbot?	Likert	Average overall satisfaction
Q6	Did the chatbot effectively resolve your problem or answer your inquiry?	Dichotomic	Percentage of inquiries effectively resolved
Q7	Have you ever left the conversation with the chatbot before receiving a satisfactory response?	Dichotomic	Percentage of users who abandoned the conversation
Q8	Would you use this chatbot again in the future?	Dichotomic	Percentage of users with the intention to reuse
Q9	How would you compare the effectiveness of the chatbot to that of a human agent?	Likert	Perception of the chatbot compared to human support
Q10	How many times did the chatbot fail to understand or process your request?	Likert/ Intervals	Average number of errors or failures per session

Source: Own Elaboration.

V. RESULTS

Table VII shows the results obtained for questions Q1, Q2, Q3, Q4, and Q5, using a scale from 1 "Very Dissatisfied" to 5 "Very Satisfied". To calculate the average score per question, the sum of the products of each response value (1 to 5) and the number of users who selected that option was divided by a total of 32 users.

TABLE VII. RESULTS OBTAINED FOR QUESTIONS USING THE LIKERT SCALE

Result	*Q1*	*Q2*	*Q3*	*Q4*	*Q5*
5 - Very satisfied	21	20	23	25	20
4 - Satisfied	11	9	9	7	11
3 - Neutral	0	3	0	0	1
2 - Dissatisfied	0	0	0	0	0
1 – Very dissatisfied	0	0	0	0	0
Average	4.66	4.53	4.72	4.78	4.59

Source: Own Elaboration.

Regarding Accuracy of the Responses (Q1), 21 participants, representing 66%, indicated that they were Very Satisfied, and 11 participants, or 34%, responded that they were Satisfied. This indicates that the chatbot provides accurate and precise responses, with an average accuracy score of 4.66. The comparison is shown in Fig. 4.

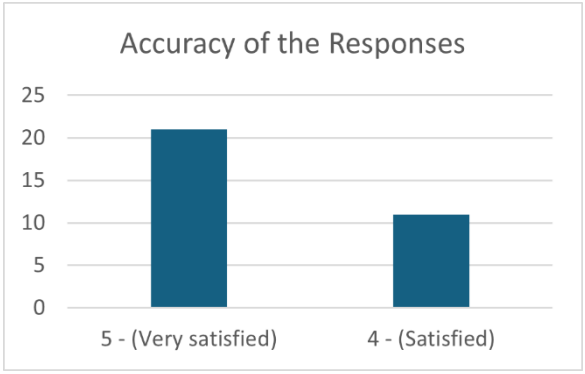


Fig. 4. Accuracy of the response results. Source: Own elaboration.

Regarding the Relevance of the Responses (Q2), 20 participants, representing 63%, indicated they were Very Satisfied; 9 participants, or 28%, responded that they were Satisfied; and 3 participants, representing 9%, expressed a Neutral opinion. This indicates that the chatbot is effectively providing relevant responses to users, with the majority expressing high satisfaction. The average frequency score was 4.53, as shown in Fig. 5.

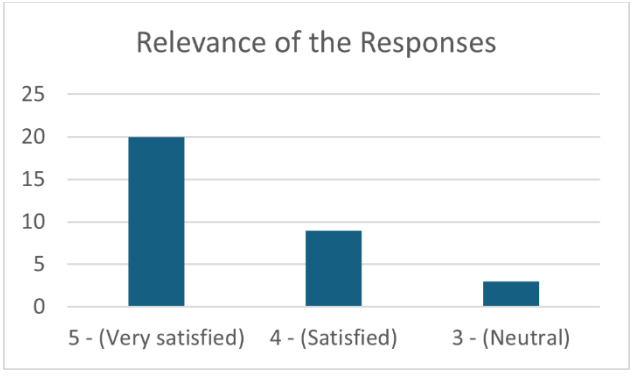


Fig. 5. Relevance of the response results. Source: Own elaboration.

Regarding Response Time (Q3), 23 participants (72%) indicated that they were Very Satisfied, and 9 participants

(28%) responded that they were Satisfied, leading to the conclusion that satisfaction with the chatbot's response time is adequate. The comparison is presented in Fig. 6.

Fig. 6. Response time results. Source: Own elaboration.

Regarding Ease of Use (Q4), it was found that 25 participants (78%) indicated they were Very Satisfied, and 7 (22%) responded that they were Satisfied. This indicates that the chatbot interface successfully provides a user-friendly tool, with an average satisfaction score of 4.78. The comparison is presented in Fig. 7.

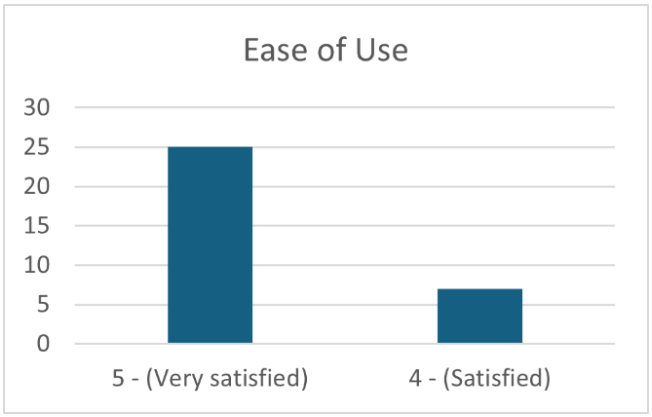


Fig. 7. Ease of use results. Source: Own elaboration.

Regarding the Overall Satisfaction Level (Q5), it was found that 20 participants (63%) indicated they were Very Satisfied, 11 (34%) responded that they were Satisfied, and 1 (3%) expressed a Neutral opinion. This indicates that the chatbot demonstrates a high level of overall satisfaction, with an average response score of 4.59. The comparison is shown in Fig. 8.

Table VIII shows the frequency of responses categorized into two options, 'Yes' and 'No', for the dichotomous questions.

Regarding the Question Resolution Rate (Q6), all participants (100%) perceived that the chatbot effectively resolved their submitted question, as shown in Fig. 9.

Regarding the Abandonment Rate (Q7), 5 participants (16%) indicated that they had abandoned the conversation before receiving a satisfactory response, while 27 participants (84%) did not abandon the conversation, resulting in a total abandonment rate of 16%. The comparison is shown in Fig. 10.

Fig. 8. Overall satisfaction level results. Source: Own elaboration.

TABLE VIII. RESULTS OBTAINED FROM DICHOTOMOUS SCALE QUESTIONS

ID	*Question*	*Yes*	*No*
Q6	Question Resolution Rate	32	0
Q7	Abandonment Rate	5	27
Q8	Intention to Reuse	32	0

Source: Own Elaboration.

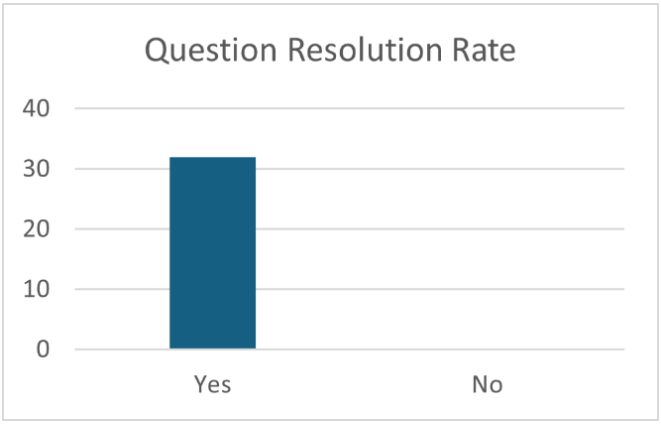


Fig. 9. Question resolution rate results. Source: Own elaboration.

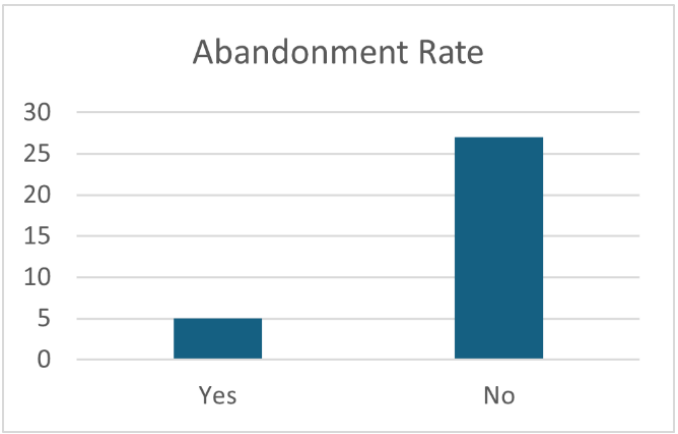


Fig. 10. Abandonment rate results. Source: Own elaboration.

Regarding the Intention to Reuse the Application (Q8), it was found that all participants (100%) would be willing to use this tool in the future. The comparison is shown in Fig. 11.

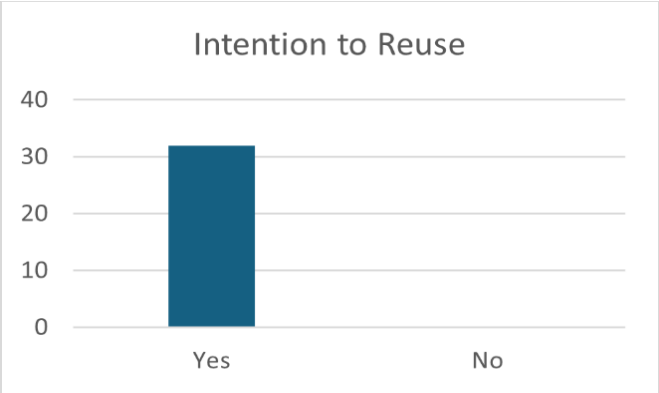


Fig. 11. Results on the intention to reuse. Source: Own elaboration.

Table IX shows the results obtained for the ninth question related to the comparison with human support, using a rating scale from "Much Worse" to "Much Better".

TABLE IX. RESULTS OBTAINED FOR QUESTION 9

Results	*Q9*
Much better	13
Better	12
Equal	7
Worse	0
Much worse	0

Source: Own Elaboration.

Regarding the Comparison with Human Support, 13 participants (41%) indicated that they consider the chatbot to be Much Better, 12 (37%) consider it Better, and 7 (22%) find it Equal to the attention provided by a human agent. The comparison is shown in Fig. 12.

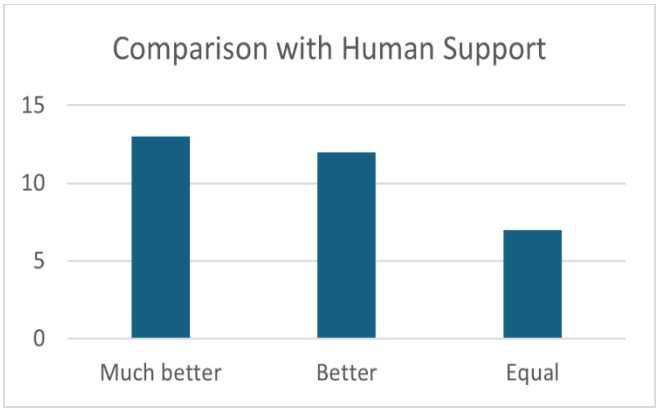


Fig. 12. Results on the comparison with human support. Source: Own elaboration.

Finally, Table X displays the results for question 10, focused on the number of errors encountered during queries.

Regarding the Error Rate (Q10), 20 participants (63%) reported not encountering any errors, 4 (13%) encountered 1 error, and 8 (25%) encountered 2 errors. There is an average of 0.63 errors found per user. The comparison is shown in Fig. 13.

TABLE X. RESULTS OBTAINED FOR Q10 REGARDING ERRORS

Results	*Q10*
0	20
1	4
2	8
3	0
4 or more	0

Source: Own Elaboration.

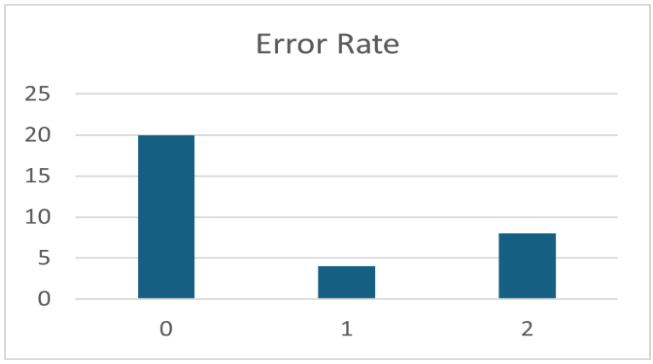


Fig. 13. Comprehension or processing errors encountered by users in the chatbot. Source: Own elaboration.

VI. DISCUSSION

The results of this study demonstrate that a chatbot integrated with artificial intelligence can be an effective solution to improve access to information about health insurance in Peru. Validation was carried out using 50 questions evaluated on the criteria of clarity, relevance, coherence, and accuracy, yielding an overall mean response accuracy of 82%, which, together with the usability findings, indicates that the chatbot achieved a high level of acceptance and reliable functionality.

In comparison with related work, study [5] employs an RNN along with a Sequential Matching Framework (SMF) that transforms words into vectors and analyzes complex relationships between expressions, achieving high precision (0.97). In contrast, the present work uses a simple RNN without SMF, focused on classifying frequently asked questions about health insurance. Instead of adding complexity, it is complemented by the OpenAI model, which effectively interprets context. This combination provides fast and clear answers to specific queries, reaching 82% compliance and 97% user satisfaction in testing.

In study [6], bidirectional LSTM neural networks (BiLSTM) are combined with a Text Convolutional Neural Network (TextCNN). Although both models rely on neural networks, the present study focuses on simple RNNs, which are more efficient for short questions and well-defined contexts such as EsSalud health insurance. This integration enables the system to respond to queries without the need for computationally intensive architecture or elaborate training processes, maintaining a good balance between performance and simplicity. In contrast, the hybrid BiLSTM-TextCNN approach focuses on recognizing general textual patterns, which are more complex.

On the other hand, study [7] employs advanced architecture that combines LSTM with a Seq2Seq model to develop highly complex chatbots, especially useful for extended dialogues. In contrast, this study opts for a simple RNN, avoiding the complexity of the Seq2Seq approach and focusing on brief, specific queries related to health insurance. This simplicity enhances efficiency while achieving similar results. Furthermore, the use of OpenAI's language model improves the understanding of question semantics and enables the generation of clearer and more appropriate responses. This makes the system more useful and practical when compared to models like DeepQA, which are more oriented toward maintaining long and complex conversations.

In study [8], the Bablibot chatbot was developed to answer questions related to immunization in Pakistan, utilizing a continuous learning process. The result was a 90% user acceptance rate. In comparison, the proposed model focuses on the domain of health insurance, an area in which users often have limited prior knowledge. Despite this, the proposed system achieved satisfaction rates of 63% (very satisfied) and 34% (satisfied), resulting in a total satisfaction score of 97%.

Additionally, study [9] employs Dempster-Shafer theory to fuse data from multiple sources, enhancing accuracy in complex and highly contextual questions. In contrast, the present work uses a simple RNN architecture without external data fusion, focusing on delivering immediate and accurate responses within the specific domain of health insurance. This narrower contextual scope enables faster response times and easier implementation.

Finally, in study [10], a variation of RNN known as Transposed GRU was used, which reverses the input data sequence during processing. In comparison, this study opted for a simple RNN to generate chatbot responses, which was likewise validated through user feedback via a survey. In the user validation, the present study achieved a perceived relevance of responses of 91% (63% "Very Satisfied" and 28% "Satisfied"), compared to the 76% obtained in the referenced study. Furthermore, the overall satisfaction level perceived by users was predominantly positive at 97% (63% "Very Satisfied" and 34% "Satisfied"), contrasting with the 76.2% acceptance rate reported in the referenced work.

Although the model demonstrated strong results, it has limitations in handling complex questions or extended dialogues due to the use of a simple RNN. In addition, depending on the OpenAI-provided AI model, constraints may arise in certain contexts. The validation was also conducted with a limited sample size, and expanding the study could provide more comprehensive feedback to identify critical improvement areas.

Despite these limitations, the model stands out for its simplicity, speed, and ease of implementation. These features make it potentially adaptable to other industries, such as education, public services, or other insurance providers, where quick and clear access to information is equally essential. Adjusting the content and context would be sufficient to address emerging community needs, thanks to the scalability of the system. This scalability depends on the databases to be incorporated for adaptation to other health domains, as well as the expansion of training questions to be used, thereby enriching the information possessed by the developed model.

VII. Conclusion

Artificial intelligence presents great potential to support customer service in the healthcare sector. Through an AI- and RNN-based chatbot, it is possible to automate repetitive tasks such as answering frequently asked questions and providing basic information about health insurance, thanks to the RNN algorithm's ability to select the most optimal response and the AI model's natural language generation capabilities. This technology not only improves response speed and reduces wait times but also demonstrates that intelligent automation can be integrated into both public and private healthcare institutions to enhance information accessibility and efficiency. Despite their relatively simple implementation, this type of solution demonstrates the feasibility of modernizing service channels using accessible and efficient tools, promoting a better experience for both insured and uninsured individuals, with the potential to enrich the chatbot's responses by increasing the training data for the RNN model or employing more specialized architectures tailored to this type of algorithm.

The evaluation results showed an overall mean response accuracy of 82% and high satisfaction levels across all criteria, indicating that the developed chatbot provides reliable, coherent, and contextually relevant responses. These findings suggest that combining an RNN classifier with a generative model like GPT-3.5 constitutes an effective hybrid strategy for managing health-related queries, achieving performance close to that of human support.

It was found that the developed chatbot and the designed interface were well received by users, with a majority of responses indicating "Very Satisfied" for Question P1 on perceived response accuracy (66%) and P5 on overall satisfaction (63%). Additionally, there was a positive perception compared to human support, with 78% of participants stating the experience was "Much Better" (41%) or "Better" (37%). Finally, the intention to reuse the tool reached 100%, meaning that all participants found the chatbot useful.

From a practical perspective, these results imply that implementing similar chatbots could significantly reduce administrative bottlenecks and improve the user experience in state health systems by facilitating immediate and accurate access to insurance information. Scientifically, the study contributes evidence that hybrid AI models can successfully merge deterministic and generative reasoning to improve the interpretability and adaptability of automated agents in specialized domains.

Furthermore, the 100% intention to reuse the tool and the 78% positive comparison to human support highlight users' trust in this type of system, reflecting its potential to complement rather than replace human staff. Future research could explore expanding the dataset used for RNN training, integrating reinforcement learning for adaptive responses, and testing the model in other public service areas, thus consolidating AI's role in citizen-oriented digital transformation.

ACKNOWLEDGMENT

Thanks to the Dirección de Investigación of the Universidad Peruana de Ciencias Aplicadas for their support to carry out this research work through the UPC-EXPOST-2025-2 incentive.

ETHICAL APPROVAL

The study involved only anonymous usability testing conducted by voluntary users. No clinical data or personal information were collected, and the study posed no risk to participants. Although formal ethical approval was not required, the research was conducted in accordance with the general ethical principles outlined in the Declaration of Helsinki.

INFORMED CONSENT

When accessing the application, all participants were clearly informed about the purpose of the study and that no personal data was required. Participants voluntarily agreed to take part in the usability testing, and upon entering the system, they enabled the option confirming their understanding of the application's purpose and provided their consent through the platform itself to proceed with its use.

DISCLOSURE STATEMENT

The authors report there are no competing interests to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in the Open Science Framework (OSF) at: https://doi.org/10.17605/OSF.IO/F49Q7.

REFERENCES

[1] Superintendencia Nacional de Salud (2022). Registro Nominal de Asegurados. REGINA Boletín Informativo. https://ipe.org.pe/wp-content/uploads/2022/02/SUSALUD-Boletin-REGINA-31-de-enero-2021.pdf

[2] Centro Nacional de Planeamiento Estratégico. (2023). Mayor población afiliada a un sistema de salud. Observatorio Ceplan. https://observatorio.ceplan.gob.pe/ficha/t24

[3] Ramos, A. (2024). Ingresos de EsSalud se duplicaron, pero consultas ambulatorias de afiliados solo crecieron 5% en 15 años - Infobae. Infobae. https://www.infobae.com/peru/2024/01/14/ingresos-de-essalud-se-duplicaron-pero-consultas-ambulatorias-de-afiliados-solo-crecieron-5-en-15-anos/

[4] Tovar, A. (2023). EsSalud: radiografía de la crisis que afecta a 12 millones de asegurados. Salud con Lupa. https://saludconlupa.com/noticias/essalud-radiografia-de-la-crisis-que-afecta-a-12-millones-de-asegurados/

[5] Nuruzzaman, M., & Hussain, O. K. (2020). IntelliBot: A Dialogue-based chatbot for the insurance industry. Knowledge-Based Systems, 196, 105810. https://doi.org/10.1016/j.knosys.2020.105810

[6] Li, Z., Yang, X., Zhou, L., Jia, H., & Li, W. (2023). Text Matching in Insurance Question-Answering Community Based on an Integrated BiLSTM-TextCNN Model Fusing Multi-Feature. Entropy, 25(4), 639. https://doi.org/10.3390/e25040639

[7] Choudhary, P., & Chauhan, S. (2023). An intelligent chatbot design and implementation model using long short-term memory with recurrent neural networks and attention mechanism. Decision Analytics Journal, 9, 100359. https://doi.org/10.1016/j.dajour.2023.100359

[8] Siddiqi, D. A., Miraj, F., Raza, H., Hussain, O. A., Munir, M., Dharma, V. K., Shah, M. T., Habib, A., & Chandir, S. (2024). Development and feasibility testing of an artificially intelligent chatbot to answer immunization-related queries of caregivers in Pakistan: A mixed-methods study. International Journal Of Medical Informatics, 181, 105288. https://doi.org/10.1016/j.ijmedinf.2023.105288

[9] Nguyen-Mau, T., Le, A., Pham, D., & Huynh, V. (2024). An information fusion based approach to context-based fine-tuning of GPT models. Information Fusion, 104, 102202. https://doi.org/10.1016/j.inffus.2023.102202

[10] Gidey, K., Beyene, H., Haileslassie, F. & Berihu, H. (2025). Apprentice bot model design and implementation for psychological clients' therapy. Discover Applied Sciences, 7(3), 170. https://doi.org/10.1007/s42452-025-06515-2

[11] Martín, A. (2018). El contrato de seguro de asistencia sanitaria y la protección del asegurado. Universidad de Almería. http://hdl.handle.net/10835/7228

[12] Gutiérrez, C., Romaní, F. R., Wong, P., & Del Carmen Sara, J. (2018). Brecha entre cobertura poblacional y prestacional en salud: un reto para la reforma de salud en el Perú. Anales de la Facultad de Medicina, 79(1), 65. https://doi.org/10.15381/anales.v79i1.14595

[13] White, L., Fishman, P., Basu, A., Crane, P. K., Larson, E. B., & Coe, N. B. (2019). Medicare expenditures attributable to dementia. Health Services Research, 54(4), 773-781. https://doi.org/10.1111/1475-6773.13134

[14] Mueller, J., Makridis, C. A., Tiffany, T., Borkowski, A. A., Zachary, J., & Alterovitz, G. (2024). From theory to practice: Harmonizing taxonomies of trustworthy AI. Health Policy OPEN, 100128. https://doi.org/10.1016/j.hpopen.2024.100128

[15] Lim, H., Park, Y., Hong, J. H., Yoo, K., & Seo, K. (2024). Use of machine learning techniques for identifying ischemic stroke instead of the rule-based methods: a nationwide population-based study. European Journal Of Medical Research, 29(1). https://doi.org/10.1186/s40001-023-01594-6

[16] Weinberg, S. T., Monsen, C., Lehmann, C. U., Leu, M. G., Webber, E. C., Alexander, G. M., Beyer, E. L., Chung, S. L., Dufendach, K. R., Hamling, A. M., Harper, M. B. & Kirkendall, E. S. (2021). Integrating Web Services/Applications to Improve Pediatric Functionalities in Electronic Health Records. Pediatrics, 148(1). https://doi.org/10.1542/peds.2021-052047

[17] Palta, R., Angwin, J. & Nelson, A. (2024). How We Tested Leading AI Models Performance on Election Queries. The AI Democracy Projects. https://www.proofnews.org/how-we-tested-leading-ai-models-performance-on-election-queries/

[18] Mohammad, M. (2023). An evaluation of the critical success factors impacting artificial intelligence implementation. International Journal of Information Management, 69, 102545. https://doi.org/10.1016/j.ijinfomgt.2022.102545

[19] Erickson, J. (2024). MySQL: qué es y cómo se usa. Oracle. https://www.oracle.com/ar/mysql/what-is-mysql/

[20] Waqas, M. & Wannasingha, U. (2024). A critical review of RNN and LSTM variants in hydrological time series predictions. MethodsX, 13, 102946. https://doi.org/10.1016/j.mex.2024.102946

[21] Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G. & Pedrycz, W. (2024). A comprehensive survey on applications of transformers for deep learning tasks. Expert Systems with Applications, 241, 122666. https://doi.org/10.1016/j.eswa.2023.122666

[22] Utomo, A., Wijaya, G., & Setiawan, Y. (2023). Implementation of Crowdfunding Web Application Using AWS Amplify Architecture with End-to-End Testing Using Playwright. Indonesian Journal Of Multidisciplinary Science, 2(11), 3888-3904. https://doi.org/10.55324/ijoms.v2i11.620

[23] Elordi, U., Unzueta, L., Goenetxea, J., Sanchez-Carballido, S., Arganda-Carreras, I., & Otaegui, O. (2020). Benchmarking Deep Neural Network Inference Performance on Serverless Environments With MLPerf. IEEE Software, 38(1), 81-87. https://doi.org/10.1109/ms.2020.3030199

[24] Morgado, J. (2024). Deploy your first web app on AWS with AWS Amplify, Lambda, DynamoDB, and API Gateway. https://dev.to/juliafmorgado/deploy-your-first-web-app-on-aws-with-aws-amplify-lambda-dynamodb-and-api-gateway-2ah7

[25] Mandic, D., Miscevic, G. & Bujišić, L. (2024). Evaluating the Quality of Responses Generated by ChatGPT. Metodicka praksa. 27(1), 5-19. https://doi.org/10.5937/metpra27-51446

APPENDIX A: PSEUDO CODE OF THE PROPOSED RNN–AI MODEL

```
System Initialization:
a. Load the pre-trained Recurrent Neural Network (RNN) model from
"modelo.keras".
b. Load the tokenizer ("tokenizer.pkl") and label encoder
("label_encoder.pkl"), required for text processing and response
interpretation.
c. Define global system parameters:
i. Maximum sequence length: 20
ii. Confidence threshold: 0.7

Query Reception:
a. The system receives the user's input (Q).
b. Validate that the question is not empty and does not contain
unrecognized or invalid characters.

Text Preprocessing and Normalization:
a. Convert the entire query to lowercase.
b. Tokenize the text, converting each word into a numeric index according
to the trained vocabulary.
c. Apply padding to the resulting sequence to reach the predefined maximum
length.

RNN Inference:
a. Feed the processed sequence into the RNN model.
b. The model produces a probability distribution (P) over all possible
responses in its trained set.

Prediction Confidence Evaluation:
a. Identify the maximum probability (prob_max = max(P)) and the associated
index (response_index = argmax(P)).
b. If the obtained probability is greater than or equal to the defined
threshold, decode the response using the label encoder.

Base Response Generation:
a. If (prob_max ≥ confidence_threshold), obtain the preliminary response
(R) from the identified index.
b. If no relevant match is found, set a generic response:
R ← "I'm sorry, I don't have that information."

Response Postprocessing:
a. Clean and format (R), removing duplicates, errors, or unnecessary
characters.

Integration with the Generative Model (OpenAI):
a. Send the base response (R) as input to OpenAI's language model.
b. OpenAI analyzes, reformulates, and enriches the content, producing a
more natural, clear, and contextualized final response: (R_final).

Error and Exception Handling:
a. If an error occurs while loading model files or performing inference,
log the event in the system logs.
b. In such cases, return the message: "The request could not be processed."

System Output:
Present the final response (R_final) to the user as the result of the
chatbot interaction.
```

Text Information Data Mining Method in Natural Language Processing Tasks

Shengguo Guo*, Dandan Xing
School of Information Engineering, Zhenzhou College of Finance and Economics, Zhengzhou, China

***Abstract*—Text mining methods often rely on a single data source or simple word frequency statistics, making it difficult to capture multi-source text semantic associations and local contextual dependencies, resulting in poor mining accuracy. Therefore, a method for text information data mining in natural language processing tasks is proposed. Using Python web crawlers to obtain multi-source text data, after preprocessing such as cleaning, segmentation, and removal of stop words, a Vector Space Model (VSM) is used for text representation, and a TF-IDF (Term Frequency Across Document Frequency) weight optimization mechanism is introduced to enhance feature semantic representation. On this basis, a semantic enhancement system is constructed based on the BERT classification model in the field of natural language processing. Through the self-attention mechanism of multi-layer Transformer encoders, semantics are aggregated to effectively capture local contextual dependencies, and context-sensitive word vectors are generated by the output layer. Finally, by fine-tuning the parameters of the BERT (Bidirectional Encoder Representation from Transformers) model and combining it with the Softmax function, precise mining of text information data categories was achieved. The experimental results show that in the embedding experiment of sports news headlines, this method can form a semantic aggregation structure with clear domain logic for word vectors; In the cross domain short text classification experiment, the overall accuracy of this method on the dataset reached 95.7%, which was 19.5% and 18.7% higher than the comparative methods, effectively solving the cross domain ambiguity problem in natural language processing.**

Keywords*—Natural language processing; text information; data mining; VSM; TF-IDF; BERT

I. Introduction

Text information data is widely distributed in scenarios such as social media, academic platforms, enterprise systems, and medical databases, covering diverse types such as online reviews, academic papers, financial annual reports, and electronic medical records. These data continue to grow at a rate of hundreds of millions per day, forming a huge unstructured information cluster. Different from structured data, text information data exhibits significant looseness and complexity: online jargon, abbreviations, and emojis are mixed in social media comments; academic literature involves professional terms and complex sentence patterns; industry jargon and ambiguous expressions exist in enterprise reports; medical records contain medical codes and colloquial records. These characteristics make it difficult for computers to directly and efficiently process and analyze text information [1]. Facing such complex and disorderly text information data, how to extract valuable knowledge from it has become an important topic in the field of data science. Studying text information data mining methods aims to break through the unstructured barriers of text and extract implicit patterns, trends, and relationships from massive texts [2]. Taking an e-commerce platform as an example, by mining consumer reviews, not only can specific pain points of products in dimensions such as function, quality, and service be identified, but also the evolution trend of user needs over time can be traced; in the field of academic research, in-depth mining of literature can reveal interdisciplinary hotspots, technology evolution paths, and key unsolved problems; in the medical field, analysis of medical record texts can discover potential associations between disease symptoms and treatment plans, and even predict the incidence patterns of diseases [3]. Through these mining efforts, disorderly text data is transformed into structured knowledge, achieving a qualitative change from "data accumulation" to "knowledge precipitation", providing a core basis for decision-making, technological innovation, and problem-solving.

In the research on text data mining in the industry, Kim, H. J. et al. proposed a conditional generative adversarial multiple imputation network based on unsupervised data mining. An auxiliary classifier was developed through a pre-training algorithm combined with implicit class labels. The sorting point recognition clustering structure algorithm of fuzzy clustering was used to learn implicit classification information, and multiple imputation was applied to the original energy dataset to improve the robustness and reliability of the imputation results [4]. However, when explicit class labels are not available, it is difficult to effectively capture the semantic associations of multi-source texts, resulting in poor accuracy of mining results. Rahman, U. et al. used two classification algorithms, Naive Bayes and Support Vector Machine, to perform binary classification on the text data generated by the regular maintenance of industrial plant equipment. The model was trained and tested by analyzing the labeled data to achieve the early classification of equipment failures, thereby improving the industry maintenance plan [5]. However, this method does not consider capturing the long-distance dependence relationships between different words in the text, which may lead to a decrease in the classification mining accuracy in complex fault scenarios. Kumar, M. R. P. et al. mined context-related subjective words in the text through a context-based sentiment dictionary, calculated the weight scores of subjective terms and complete reviews, and then classified the text sentiment with the help of a bidirectional LSTM model. At the same time, an oversampling technique was used to generate new synthetic text samples using semantic information to solve the

* Corresponding author.

data imbalance problem and achieve sentiment classification mining of imbalanced corpora [6]. However, the context-based sentiment dictionary needs to be manually defined or pre-annotated, and it is difficult to cover new sentiment expressions in dynamic contexts, resulting in poor capture of the semantic associations of multi-source texts and poor adaptability to out-of-domain texts. Kim, H. et al. proposed a flexible periodic pattern data mining method for incremental time series databases, which processes data streams in real time by enhancing the data structure and mines periodic patterns containing uncertain events, providing decision support for scenarios such as oil price fluctuation prediction and traffic congestion analysis [7]. However, this method cannot accurately express the semantic differences of words in different contexts, easily leading to insufficient accuracy of text representation and poor mining performance.

In summary, existing text information data mining methods still have significant limitations in deep semantic modeling and cross-domain generalization ability: most models fail to effectively integrate the statistical features of text with global contextual semantics, resulting in insufficient semantic understanding of polysemous words, domain-specific terminology, and cross-domain short texts, thereby restricting further improvement in mining accuracy and practicality. Natural Language Processing (NLP), as the core technology for processing text data, provides crucial support for text information data mining [8]. In recent years, NLP technologies driven by deep learning, such as pre-trained language models like BERT and GPT, have significantly improved the ability to understand text semantics through learning from massive corpora, greatly promoting the development of text information data mining [9]. These technologies can accurately capture the sentiment tendency, entity relationships, and topic context in text, demonstrating great value in practical applications: in the e-commerce field, mining the semantic features of consumer reviews can assist in product optimization; in the medical industry, mining medical record texts helps with clinical decision-making; in academic research, literature semantic analysis can help keep up with the cutting-edge of the discipline [10]. Therefore, in view of the limitations of the above methods, this study focuses on the research of text information data mining methods in natural language processing tasks, in order to promote the cross-innovation of natural language processing and data mining technologies, and provide efficient transformation solutions from text data to structured knowledge for various industries. The specific technical route is as follows:

- Firstly, a Python web crawler is used to construct a multi-source heterogeneous text corpus, and strict cleaning, segmentation, and stop word preprocessing are performed to provide a high-quality, cross-domain training and evaluation foundation for the model, overcoming the bias problem caused by a single data source.
- Then, based on the vector space model representation, the TF-IDF weight optimization mechanism is introduced. This step goes beyond simple word frequency statistics and can enhance key semantic features with high discrimination, providing strong feature inputs rich in statistical information for subsequent deep models and improving the discriminative power of semantic representation from the perspective of feature engineering.
- Subsequently, a semantic enhancement system based on the BERT classification model was constructed, which utilizes the self-attention mechanism of the Transformer encoder to dynamically aggregate global contextual information, accurately capturing long-distance dependencies and polysemy between words, fundamentally solving the core deficiency of weak contextual modeling ability in traditional models.
- Finally, the optimized TF-IDF features are deeply fused with the context vector of BERT through fine-tuning strategies, and precise classification is achieved using the Softmax function. This fusion scheme not only inherits the advantages of traditional methods with clear features, but also has the semantic understanding ability of deep models.
- Design a cross-domain short text classification experiment to verify the comprehensive performance of the proposed framework in handling polysemous words and domain-specific terms from dimensions such as semantic representation quality, computational efficiency, and term robustness on a multi-domain news headline dataset, including sports, entertainment, and real estate.

II. Optimize Text Feature Vectors Based on TF-IDF

In the text information data mining task in the field of Natural Language Processing (NLP), the core is to convert unstructured text into computable semantic information, and data acquisition and preprocessing are the basic links to achieve this goal [11]. This study applies web crawler technology based on Python to crawl the required text data from various text data sources, such as social media, news texts, and academic literature. Collect information such as text content and publication time, and store the data in text format to provide the original corpus for subsequent text information data mining [12].

A. Text Information Data Preprocessing

The preprocessing process of text information data is shown in Fig. 1, and the specific process is as follows:

1) Text data cleaning: Remove the possible special symbols and extra blanks in the crawled text data. Extra blanks may separate complete words, resulting in incorrect word segmentation, reducing the word segmentation accuracy, and thus affecting the final text information data mining effect. Therefore, it is necessary to ensure the standardization of text data through cleaning [13].

2) Chinese text word segmentation: Select the jieba word segmentation library based on the Chinese word library in Python to perform word segmentation on the text data. Split the text content into characters, words, or phrases so that the computer can accurately recognize the words in the text. This

method has high word segmentation accuracy and fast speed, which can improve the accuracy of subsequent text classification, clustering, and other tasks [14].

3) Remove obsolete terms: Compare with the obsolete terms library to filter the text data, delete words such as modal particles, prepositions, and pronouns that have no actual meaning and may affect the text mining results, increase the keyword density of the text, and at the same time reduce the feature dimension, thereby improving the accuracy and efficiency of text mining [15].

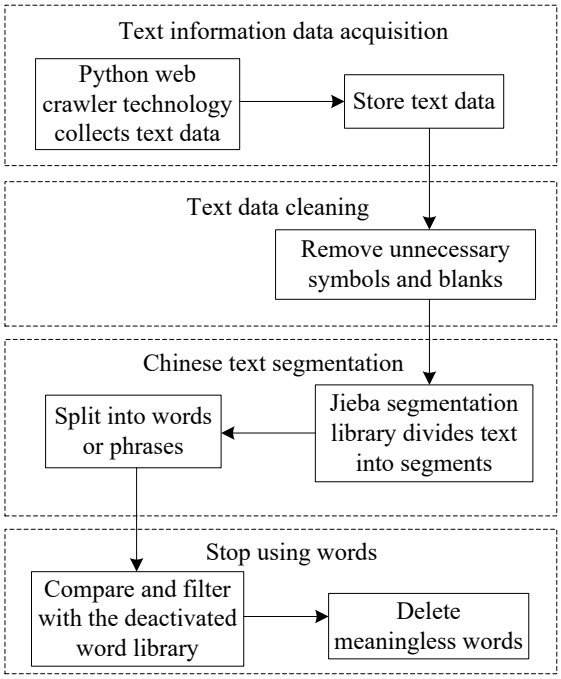


Fig. 1. Preprocessing process of text information data.

B. Text Representation Based on Vector Space Model

Text representation, as a key bridge connecting natural language and machine understanding, directly determines the accuracy and efficiency of subsequent mining tasks. After acquiring and preprocessing text data, it is necessary to further transform unstructured text into a structured representation that can be processed by a computer, laying a foundation for in-depth semantic analysis. There are mainly three common text representation models: the Boolean model, the probabilistic model, and the vector space model. The vector space model (VSM), as one of the most widely used and effective methods currently, has its technical advantages directly reflected in the processing efficiency and mining accuracy of actual text data [16]. The smallest data unit in the vector space model is the feature term, and words, phrases, and word groups can all be used as feature terms for processing. Regarding the text c as a n -dimensional vector in the vector space, as shown in Eq. (1):

$$c=\left(\left(\varepsilon_1,\omega_1\right),\left(\varepsilon_2,\omega_2\right),\cdots,\left(\varepsilon_n,\omega_n\right)\right) \tag{1}$$

where, ε_i represents the i -th feature term of the text c; ω_i represents the feature weight corresponding to the i -th feature term of the text c, and the magnitude of the feature weight indicates the amount of text category information contained in this feature.

C. Weight Optimization Based on TF-IDF

The text data representation based on the vector space model realizes the structured processing of text through the vector mapping of feature terms and weights. However, this method relies on word frequency statistics and lacks semantic association modeling. Therefore, the TF-IDF (Term Frequency-Inverse Document Frequency) weight optimization representation method is introduced to enhance the semantic representation ability on the basis of retaining the VSM feature weight system and adapting to the complex semantic requirements in text information data mining [17]. For example, in VSM, "computer" and "PC" are regarded as independent dimensions with no associated weights, while TF-IDF ensures that both obtain high weights in scientific and technological texts through IDF, indirectly improving the similarity of semantically related texts.

For the text collection $C=\left\{c_1,c_2,\cdots,c_M\right\}$, M is the total number of texts in the text collection C. Following the feature term processing logic of VSM, the feature term set $\left\{\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n\right\}$ is obtained. Taking each feature term ε_i as a dimension of the vector space, the weight ω_i of each feature term ε_i in the document is calculated through TF-IDF, as shown in Eq. (2) to Eq. (4):

$$\omega_i=\omega\left(\varepsilon_i,C_j\right)=\frac{TF\left(\varepsilon_i,C_j\right)\times IDF\left(\varepsilon_i\right)}{\sqrt{\sum_{\varepsilon_i\in C_j}\left[TF\left(\varepsilon_i,C_j\right)\times IDF\left(\varepsilon_i\right)\right]^2}} \tag{2}$$

$$TF\left(\varepsilon_i,C_j\right)=\frac{D\left(\varepsilon_i,C_j\right)}{D\left(C_j\right)} \tag{3}$$

$$IDF\left(\varepsilon_i\right)=\log\left(\frac{M}{D\left(\varepsilon_i\right)}+0.01\right) \tag{4}$$

where, $TF\left(\varepsilon_i,C_j\right)$ represents the relative frequency of the feature term ε_i in the single text C_j; $IDF\left(\varepsilon_i\right)$ is the inverse document frequency, which measures the discrimination ability of the feature term ε_i in the entire text collection. If the feature term ε_i appears in more texts ($D\left(\varepsilon_i\right)$ is larger), then its discrimination degree is lower. On the contrary, if it only appears in a few texts, the higher the IDF value, the greater the weight; $D\left(\varepsilon_i,C_j\right)$ is the absolute number of occurrences of the feature term ε_i in the text C_j; $D\left(C_j\right)$ is the total

number of feature terms in the text C_j; $D(\varepsilon_i)$ is the number of texts containing ε_i.

Assume that the text $C_j = (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_t)$ contains t feature terms. Multiply the weight of each feature term to the corresponding vector dimension to obtain the weighted text feature vector y_i, as shown in Eq. (5):

$$y_i = \varepsilon_i \times \omega(\varepsilon_i, C_j) \tag{5}$$

The feature vectors of each text in the text collection C is shown in Eq. (6):

$$\phi_i = \frac{\sum_{j=1}^{t} y_j}{|C_i|} \tag{6}$$

where, $\sum_{j=1}^{t} y_j$ is the semantic contribution that integrates all features within the text. For example, in scientific and technological texts, the vector weights of words such as "artificial intelligence" and "big data" will be accumulated. $|C_i|$ is the text length normalization factor.

III. Text Information Data Mining Based on Semantic Enhancement of BERT Classification Model

After optimizing the text feature vector based on TF-IDF, the BERT classification model (Bidirectional Encoder Representations from Transformers) is further introduced to achieve the upgrade from statistical representation to semantic representation, thereby capturing the long-distance dependency relationships between different words in the text and automatically learning the importance weights between words, so as to generate more accurate and context-sensitive word vectors [18]. This model adopts a bidirectional variant of the encoder-decoder architecture and consists of an input layer, an encoding layer, and an output layer. The model structure is shown in Fig. 2. The design of this model structure directly serves the classification task in text information data mining:

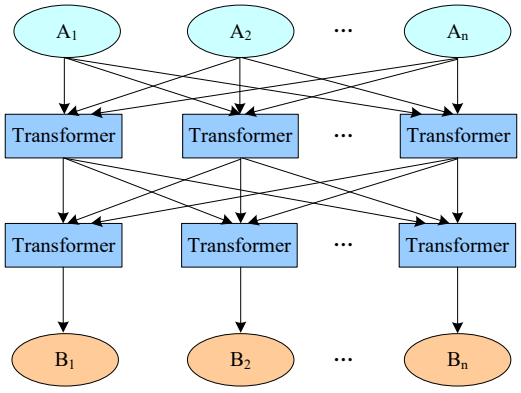


Fig. 2. Structure of BERT classification model.

The input layer combines the TF-IDF text feature vector representation $\{\phi_1, \phi_2, \cdots, \phi_n\}$ with the original text information to generate a composite text input sequence $\{A | A_1, A_2, \cdots, A_n\}$; the encoding layer is stacked by multiple layers of Transformer encoders, and each layer realizes semantic aggregation through the self-attention mechanism and the feed-forward neural network; the output layer $\{B | B_1, B_2, \cdots, B_n\}$ represents the generated word vector. A_i corresponds to B_i, and i is the i -th word of the input text.

Under the self-attention mechanism, the input vector is characterized by three vectors: the query vector Q, the key vector K, and the value vector V. The semantic association weights between words are dynamically calculated through triple operations [19], and the context of the TF-IDF feature vector representation is corrected. For example, "bank" generates different encodings in the contexts of "financial bank" and "riverbank". Input vector calculation expressions are shown in Eq. (7) to Eq. (9):

$$Q = f(q(A)) \tag{7}$$

$$K = f(k(A)) \tag{8}$$

$$V = f(v(A)) \tag{9}$$

In the equation, $f(.)$ is the linear transformation function of *linear*.

During the calculation process, each word vector participates with other vectors. Therefore, in the specific algorithm design, it is necessary to construct the corresponding weight matrix and continuously learn and update it during training. When encoding a word, not only the current word but also the context in which the word is located must be considered, so that the current word vector is integrated with the overall context, which is the self-attention mechanism. The query vector Q of each word will calculate scores with each key vector K in the entire sequence. The larger the inner product of Q and K, the closer the correlation. Based on the score redistribution feature, the scores are converted into probabilities through the Softmax function.

Different attention heads of the multi-head attention mechanism extract multiple groups of features to capture multi-dimensional semantic associations of the text, which directly serve the category discrimination task in text information data mining [20]. The calculation of the multi-head attention mechanism is shown in Eq. (10) and Eq. (11), and the attention results of each head are generated through the equation. Considering that the Softmax function may cause the gradient vanishing problem under high-value inputs, Transformer introduces a "scaling" mechanism, using the dimension parameter u_k to normalize QK^{T} to ensure the stability of gradient optimization when fusing TF -IDF features.

$$H_i = Z\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{10}$$

$$Z(Q,K,V) = S\left(\frac{QK^{\mathrm{T}}}{H_i\sqrt{u_k}}\right)V \tag{11}$$

where, QW_i^Q is the product of the query matrix Q and the weight matrix W_i^Q of the i -th head, which is used to generate the query vector of the i -th head; KW_i^K is the product of the key matrix K and the weight matrix W_i^K of the i -th head, which is used to generate the key vector of the i -th head; VW_i^V is the product of the value matrix V and the weight matrix W_i^V of the i -th head, which is used to generate the value vector of the i -th head; H is the attention head; $Z(.)$ is the attention mechanism; $S(.)$ is the Softmax function.

The diagram of the multi-head attention mechanism is shown in Fig. 3.

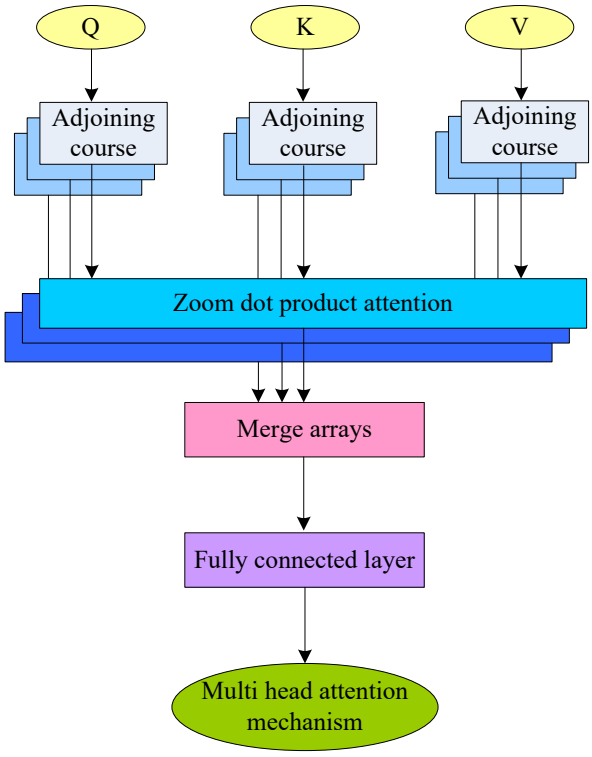


Fig. 3. Multi-head attention structure diagram.

In the model fine-tuning stage, not only the powerful ability of BERT to extract context information is retained, but also the association between TF-IDF features and class labels is further strengthened by optimizing the parameters of the classification layer [21]. The fine-tuning process minimizes the difference between the predicted probability and the true label through the cross-entropy loss function. Its core goal is to enable the BERT model to learn the discriminative representation of TF-IDF feature vectors in specific classification tasks while retaining the pre-trained semantic knowledge. Fine-tune the cross-entropy loss function used, as shown in Eq. (12):

$$L = -\frac{1}{N}\sum_{i=1}^{N}\left[x_i \times \log(\rho_i) + (1-x_i)\times\log(1-\rho_i)\right] \tag{12}$$

where, N is the total number of text samples participating in the model fine-tuning, x_i is the actual category of the i -th text sample, and ρ_i is the probability predicted for the i -th text sample.

During the fine-tuning process, by dynamically adjusting hyperparameters such as the learning rate, batch size, and number of training epochs, the model's ability to utilize TF-IDF features and capture context semantics can be effectively balanced. For example, in scientific and technological text classification, a larger learning rate can accelerate the model's response to high-weight domain words in TF-IDF (such as "deep learning"), while a smaller learning rate helps to fine-tune context relationships (such as the "although...but..." structure).

The fine-tuned model maps the feature vectors output by BERT to a class probability distribution through the Softmax function, as shown in Eq. (13):

$$S(Z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{13}$$

where, Z_i is the i -th element of the input vector of the Softmax function.

In the text information data mining framework, the input Z_i of the Softmax function is directly associated with the output of the multi-head attention mechanism: the scores Z_i corresponding to each category are the projections of the semantic representations of all words (i.e., the weighted value vectors of multi-head attention) on that category. Through Softmax calculation, the model can dynamically focus on the TF-IDF features most relevant to the category (such as high-weight words like "algorithm" and "model" in technology texts), while suppressing irrelevant information, thus achieving accurate discrimination of text categories. For example, when the input is "The application of artificial intelligence in natural language processing", the Softmax function will strengthen the influence of the TF-IDF weights of "artificial intelligence" and "natural language processing" on the classification decision, and weaken the interference of general words such as "of" and "in".

IV. Experimental Analysis

A. Experimental Settings

The experimental object of this study is the short text corpus of news titles crawled from news websites through web crawlers, covering five categories: sports, entertainment, real estate, health, and automotive. 4000 titles are collected for each category, with a total of 20000 samples. Due to the highly imbalanced samples in the dataset, the 5-fold cross-

validation method is adopted in the classification result comparison experiment: the dataset is evenly divided into five parts, and each time four parts are taken as the training set and one part as the test set for cyclic testing. Finally, the contingency is eliminated by calculating the average value of multiple test results, and at the same time, it is ensured that there is no data intersection between the training set and the test set. All texts have been cleaned and de-duplicated, segmented by jieba, and stop words removed based on the domain-specific word list to verify the performance of the text information data mining method in this study for short text classification tasks. The parameters of the experimental BERT classification model are shown in Table I. The experimental environment is shown in Fig. 4.

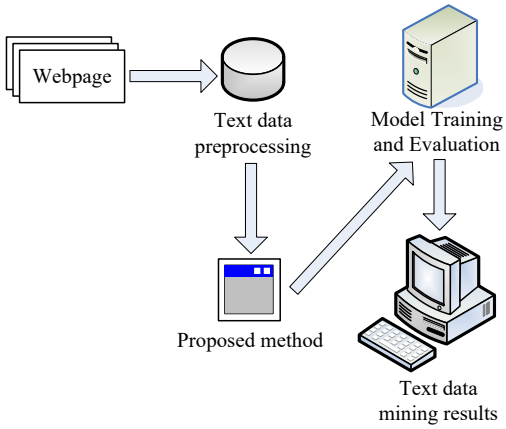


Fig. 4. Experimental environment.

TABLE I. EXPERIMENTAL BERT CLASSIFICATION MODEL PARAMETERS

Training parameters	Numerical value
Encoder layers	12
The number of self-attention heads	12
Batch sample size	15
Learning rate	0.005
number of iterations	10
Optimizer type	Adam

To verify the effectiveness of the word embedding generation method in this study for text information data, 50 professional words are systematically selected from the sports news title dataset, covering sub-domain terms such as basketball (e.g., three-point shot, dunk), football (e.g., offside, free kick), and comprehensive events (e.g., championship, home court advantage), etc.

B. Analysis of the Effectiveness of the Proposed Method

The text representation method based on semantic enhancement of the BERT classification model in this study is adopted to generate word embedding vectors through a three-layer processing process of "TF-IDF weight pre-optimization-bidirectional Transformer encoding-domain feature dynamic mapping", and the semantic association characteristics are presented in multiple dimensions by means of visualization technology. The experimental results are shown in Fig. 5.

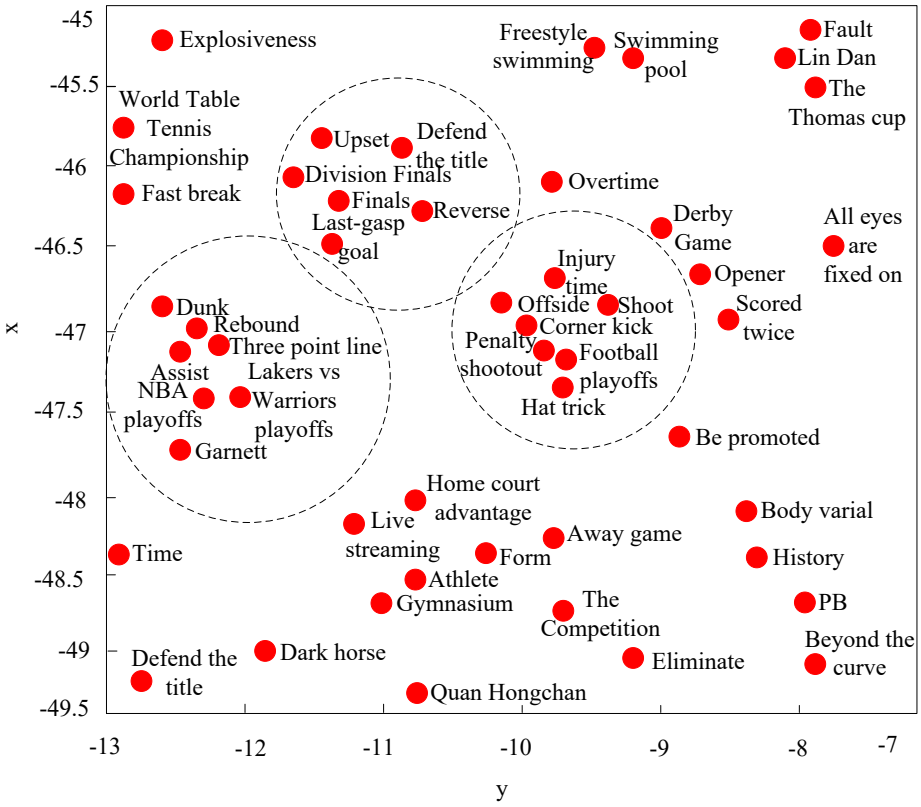


Fig. 5. Semantic spatial distribution of term embeddings in sports news texts.

Fig. 5 shows the semantic spatial distribution of the word embedding vectors generated after adopting the text representation method based on semantic enhancement of the BERT classification model in this study in the sports news title data. Several significant characteristics can be observed from the figure:

Dense clustered distribution of basketball technical terms: Basketball technical terms centered around "rebound", "assist", "three-point line", etc. show a dense clustered distribution in space. This is due to the fact that the method in this study captures the shared semantic features of "basketball movement technical actions" through the self-attention mechanism, reflecting the strong semantic associations among them, making the word vectors of similar terms highly adjacent in space.

Independent semantic clusters of football rule terms: Different from basketball technical terms, football rule terms such as "offside", "corner kick", and "penalty shootout" form independent semantic clusters. The method in this study dynamically adjusts the domain feature weights, keeping the football term clusters at a distinguishable spatial distance from the basketball technical clusters, which reflects the knowledge boundaries of different sports. For example, "offside" is a unique rule in football, and its word vector is relatively far from the word vectors of basketball terms in space, reflecting the semantic differences between them.

Semantic connections of general concepts: General concepts such as "athlete", "stadium", and "event live broadcast" on the periphery form semantic connections with the core terms through the BERT self-attention mechanism of the method in this study. For example, the vector of "stadium" has a significantly higher association strength with "home court advantage" than with "away game", which shows that the method in this study can capture the semantic changes of words in different contexts.

Hierarchical semantic network: In the word vector representation of "NBA playoffs", the method in this study strengthens the domain weights of "NBA" and "playoffs" through TF-IDF, and uses the Transformer encoder to capture the event stage features in contexts such as "Lakers vs. Warriors playoffs". This makes the word vector form a semantic chain with terms such as "finals" and "conference finals" in space, while maintaining an obvious separation from cross-domain expressions such as "football playoffs".

It can be seen that the method in this study can provide a vector representation basis with both semantic accuracy and domain adaptability for mining tasks such as intelligent classification of sports news and hot event tracking.

C. Experimental Comparative Analysis

To verify the effectiveness of the text information data mining method in this natural language processing task, multi-dimensional comparative experiments are designed: The method in this study is compared with traditional methods such as TextCNN, BiLSTM, and recently proposed Transformer baseline models RoBERTa and DistilBERT in four indicators such as semantic representation and computational efficiency. The comparison methods and configurations, and the evaluation indicator system are shown in Table II and Table III.

TABLE II. COMPARISON METHODS AND CONFIGURATION

Method	Core architecture	Optimization points for short texts
Proposed method	BERT+TF-IDF+Domain Weight Dynamic Mapping	① Strengthen domain term masking learning (such as "volume ratio"); ② Dynamically adjusting TF-IDF weights based on title word frequency
TextCNN	Convolution+Pooling+Fully Connected	① Set the window size to [1,2,3] to accommodate short text; ② Add word vectors and mixed input of word vectors
BiLSTM	Bidirectional LSTM+attention mechanism	① Hidden layer dimension 256, suitable for short sequence modeling; ② Attention weight focuses on the core words of the title (such as "new energy")
RoBERTa	Transformer+Dynamic Masking Pretraining	① Remove the next sentence prediction task; ② Train with larger batches to enhance semantic modeling capability
DistilBERT	Transformer Knowledge Distillation	① Number of parameters reduced by 40%; ② Inference speed improved, suitable for resource-constrained scenarios

TABLE III. EVALUATION INDICATOR SYSTEM

Dimension	Definition of indicators	Calculation method
Semantic representation	Cosine similarity of similar title vectors	Randomly select 100 sets of similar titles and calculate the cosine similarity mean of BERT sentence vectors
Computing efficiency	Single title inference time (ms)	The average inference time of 5000 titles in the test set
Domain term robustness	Accuracy of low-frequency term representation (less than 10 occurrences)	Calculate the cosine similarity between the vectors of low-frequency words such as "charging piles" and "science and technology innovation board" and the domain benchmark vectors
Interpretability	Title Keywords Attention Weight Contribution Rate	Using LIME algorithm to analyze the contribution of top 3 keywords (such as "school district housing" in real estate titles)

The method proposed in this study and the comparative methods both use the same BERT-base-Chinese pre-trained model as the basic encoder, focusing on issues such as domain differences, term representation, and computational efficiency of short news headlines, providing a multi-dimensional verification basis for the effectiveness of the method proposed in this study in short text data mining. The experimental results of each method in terms of semantic representation dimension, computational efficiency, domain term robustness, interpretability, etc. are shown in Table IV.

TABLE IV. COMPARISON OF MULTIDIMENSIONAL EVALUATION RESULTS OF DATA MINING METHODS FOR NEWS HEADLINES AND SHORT TEXTS

Evaluation dimensions	Proposed method	TextCNN	BiLSTM	RoBERTa	DistilBERT
Semantic representation (cosine similarity)	0.96	0.73	0.68	0.92	0.91
Calculation efficiency (single inference time/ms)	8.2	10.5	13.1	9.8	8.5
Robustness of Domain Terminology (Accuracy of Low Frequency Words)	0.97	0.59	0.65	0.93	0.90
Interpretability (contribution rate of keyword weight)	0.79	0.42	0.51	0.75	0.72

As shown in Table IV, the method proposed in this study demonstrates significant advantages in the task of mining news headline short text data. In terms of semantic representation, this method combines BERT dynamic encoding with TF-IDF weight optimization, resulting in a cosine similarity of 0.96 for similar title vectors, which is superior to RoBERTa (0.92) and DistilBERT (0.91), significantly higher than TextCNN's 0.73 and BiLSTM's 0.68. This indicates that the method proposed in this study can more accurately capture the semantic commonalities of sports titles such as "ultimate" and "comeback". In terms of computational efficiency, the proposed method only takes 8.2ms for single inference, which is superior to RoBERTa (9.8) and DistilBERT (8.5), and significantly lower than TextCNN's 10.5ms and BiLSTM's 13.1ms. This indicates that the proposed method achieves a breakthrough in efficiency while ensuring the ability of complex semantic modeling. In terms of robustness of domain terms, the method proposed in this study achieved an accuracy of 0.97 in characterizing low-frequency terms such as "plot ratio" and "new energy vehicles", surpassing RoBERTa's 0.93, DistilleBERT's 0.90, TextCNN's 0.59, and BiLSTM's 0.65. This effectively solves the problem of sparse data in short texts, making our method more robust in handling low-frequency vocabulary. In terms of interpretability, through the visualization of the attention mechanism, this method achieves a weight contribution rate of 0.79 for keywords (such as "school district housing") in classification decision-making, slightly higher than RoBERTa (0.75) and DistilBERT (0.72), significantly higher than TextCNN's 0.42 and BiLSTM's 0.51. This achieves intuitive quantification of semantic logic and improves the interpretability of the model. Based on a comprehensive multidimensional evaluation, the method proposed in this study outperforms the comparative methods in terms of semantic understanding, cross-domain generalization, and interpretability of news headlines, providing a more efficient and accurate solution for text information data mining.

To explore the application effect of the text information data mining method in this study in cross-domain short-text classification, for 5 categories (4000 items in each category) of news headline datasets, the conditional generative adversarial multiple imputation classification method in Reference [4], the binary classification method based on Naive Bayes and SVM in Reference [5] are compared with the method in this study. Through 5-fold cross-validation, starting from the number of correctly classified items in each category and typical cases, the processing ability of different methods for cross-domain ambiguous headlines is analyzed. The specific data mining results are shown in Table V.

TABLE V. COMPARISON OF TEXT INFORMATION DATA MINING RESULTS UNDER DIFFERENT METHODS

Data category	Reference[4] method	Reference[5] method	Proposed method
Sports	Correct classification number: 3120/4000 Typical case: Correct: "United Serbia final schedule" Error: "Stars participating in football live broadcasts"(Mistakenly classified as entertainment)	Correct classification number: 3050/4000 Typical case: Correct: "Player Transfer News" Error: "Artist Basketball Challenge" (Still misclassified as entertainment)	Correct classification number: 3880/4000 Typical cases: Correct: "Team training dynamics", "Celebrity basketball charity game" (Charity competitions are easily mistaken for entertainment)
Entertainment	Correct classification number: 3050/4000 Typical case: Correct: "Singer's new album released" Error: "Athletes participating in variety shows" (Mistakenly classified as sports)	Correct classification number: 3080/4000 Typical case: Correct: "Actor red carpet styling" Error: "Player attending award ceremony" (Still misjudged as sports)	Correct classification number: 3790/4000 Typical case: Correct: "Artists cross-border filming", "Athletes' variety show debut" (Athletes are easily mistaken for sports)
Real estate category	Correct classification number: 2980/4000 Typical case: Correct: "Price trend of school district housing" Error: "New policy for car down payment" (Misjudged as car)	Correct classification number: 2990/4000 Typical case: Correct: "Second hand house transaction process" Error: "Selling price of parking spaces in residential areas " (Still misjudged as a car)	Correct classification number: 3820/4000 Typical cases: Correct:"Real estate loan policy", "The selling price of the underground garage in the real estate project" (Garage is easily mistaken for a car)
Health category	Correct classification number: 3010/4000 Typical case:	Correct classification number: 3140/4000	Correct classification number: 3890/4000

	Correct: "Recommended health recipes" Error: " Air quality inspection inside the car " (Mistakenly identified as a car)	Typical case: Correct: "Traditional Chinese Medicine Health Knowledge" Error: "Methods for relieving driving fatigue " (Still mistakenly classified as a car)	Typical cases: Correct:"Healthy Eating Guide", "Healthy sitting posture of drivers" (Drivers are easily mistaken for cars)
Automobile	Correct classification number: 3090/4000 Typical case: Correct: "New car launch" Error: "Automobile property registration " (Misjudged as real estate)	Correct classification number: 3120/4000 Typical case: Correct: "Car fuel consumption test" Error: "Car transaction transfer process" (still misjudged as real estate)	Correct classification number: 3910/4000 Typical cases: Correct: "Autonomous driving technology", "Location selection for automobile 4S stores" (Site selection' is easily mistaken for real estate)

It can be seen from Table V that the classification performance of the method in this study on the 5-category (4000 items in each category) news headline datasets is significantly better than the conditional generative adversarial multiple imputation classification method in Reference [4] and the recursive-deep convolutional neural network classification method in Reference [5]. In terms of the number of correctly classified items in each category, the method in this study achieves 3880, 3790, 3820, 3890, and 3910 correctly classified items in sports, entertainment, real estate, health, and automobile categories respectively, with an overall accuracy rate of 95.7%, which is 19.5% higher than that of the conditional generative adversarial multiple imputation classification method in Reference [4] (76.2%) and 18.7% higher than that of the binary classification method based on Naive Bayes and SVM in Reference [5] (77.0%). Among them, the difference in the processing effect of cross-domain ambiguous headlines is particularly significant. For example, a "charity game" is easily misjudged as an entertainment category because of the participation of entertainment stars and the activity form leaning towards variety show interaction. But the method in this study captures the core sports attribute of "basketball" and combines features such as event rules and competitive background to avoid misjudgment, only because of the "charity" label. In "healthy sitting posture for drivers", "drivers" points to the car usage scenario and is easily classified as the "automobile" category, but the method in this study captures the core demand of "healthy sitting posture", associates features in the health field such as ergonomics and occupational disease prevention, rather than the car function itself, and classifies it as the "health category", realizing accurate text information data mining. These typical cases fully prove that the method in this study effectively solves the cross-domain ambiguity problem by virtue of the dynamic matching of domain feature weights and the context semantic association mechanism, and has significant advantages in text classification containing polysemous keywords.

V. Conclusion

This study proposes a deep mining framework that integrates TF-IDF weight optimization and BERT semantic enhancement to address the problem of insufficient semantic understanding in existing text mining methods for cross-domain short text classification. The main contributions of this study are reflected in three aspects: at the theoretical level, it explores the fusion mechanism of traditional statistical features and deep pre trained language models, providing new ideas for solving the collaborative modeling problem of statistical information and contextual semantics in text. At the methodological level, an end-to-end classification model was constructed, which enhanced the discrimination of key terms through TF-IDF and utilized BERT's self-attention mechanism to capture global context, effectively improving the model's semantic discrimination ability. On an empirical level, experiments on cross-domain news headline datasets have shown that this method outperforms various baseline models in accuracy, robustness, and interpretability, providing an effective technical solution for solving cross-domain ambiguity problems in natural language processing.

The method proposed by this research institute has significant practical application value. In scenarios such as automatic tagging of news content, monitoring of social media public opinion, and fine-grained classification of e-commerce comments, this method can accurately identify easily confused cross-disciplinary information (such as distinguishing between sports and entertainment attributes of "football stars attending charity parties"), significantly reducing the misjudgment rate. Its efficient processing capability (with a single title inference time of only 8.2ms) also enables it to adapt to online services that require high real-time performance, providing powerful tools for enterprises and platforms to achieve automated and intelligent text information processing and knowledge extraction, and has broad engineering application prospects.

Although this study has achieved the expected results, there are still some limitations. Firstly, the performance of the model depends to some extent on the matching degree between the pre-trained corpus and the target domain. When facing highly specialized or technically novel vertical domains (such as academic literature in specific disciplines), the performance may experience some attenuation. Secondly, the experiments in this study mainly focus on short texts (news headlines). Although short texts are a common form of information distribution, the method's ability to capture cross-paragraph semantic dependencies in long documents (such as research reports and lengthy comments) has not been fully validated. In addition, although the interpretability of the model is reflected through attention weights, a systematic and end-user-oriented decision interpretation reporting mechanism has not yet been formed.

Based on the results and limitations of this study, future work can be carried out in the following directions: exploring the introduction of domain adaptation techniques to enable the model to quickly migrate to professional fields such as finance and healthcare using a small number of annotated samples, improving its generalization ability and practical scope. The study combines the current model with mechanisms such as text summarization, and hierarchical attention to efficiently

process long documents and capture their core semantic and discourse structure information. Integrate cutting-edge interpretable artificial intelligence tools such as LIME and SHAP, develop visual decision path analysis functions, and enhance the transparency and credibility of models in high-risk application scenarios.

ACKNOWLEDGMENT

This work was supported by the 2025 Soft Science Research Plan Project of Henan Province "Research on the Optimization Path of New Quality Productive Forces Empowering Enterprise-led Industry-University-Research Collaborative Innovation" (Grant No. 252400410121).

REFERENCES

[1] Z. Wang, K. Ezukwoke, A. Hoayek, M. Batton-Hubert, and X. Boucher, "Correction: natural language processing (nlp) and association rules (ar)-based knowledge extraction for intelligent fault analysis: A case study in semiconductor industry," J. Intell. Manuf., vol. 36, no. 1, pp. 357-372, January 2025.

[2] R. H. Assaad, M. Mohammadi, and G. Assaf, "Determining critical cascading effects of flooding events on transportation infrastructure using data mining algorithms," J. Infrastr. Syst., vol. 30, no. 3, pp. 1.1-1.17, April 2024.

[3] M. Nazari, H. Emami, R. Rabiei, A. Hosseini, and S. Rahmatizadeh, "Detection of cardiovascular diseases using data mining approaches: application of an ensemble-based model," Cogn. Comput., vol. 16, no. 5, pp. 2264-2278, May 2024.

[4] H. J. Kim and M. K. Kim, "An unsupervised data-mining and generative-based multiple missing data imputation network for energy dataset," IEEE Trans. Ind. Inform., vol. 20, no. 11, pp. 13429-13440, November 2024.

[5] U. Rahman and M. U. Mahbub, "Application of classification models on maintenance records through text mining approach in industrial environment," J. Qual. Maint. Eng., vol. 29, no. 1, pp. 203-219, March 2023.

[6] M. R. Pavan Kumar and P. Jayagopal, "Context-sensitive lexicon for imbalanced text sentiment classification using bidirectional lstm," J. Intell. Manuf., vol. 34, no. 5, pp. 2123-2132, June 2023.

[7] H. Kim, H. Kim, S. Kim, H. Kim, M. Cho, B. Vo, J. C. W. Lin, and U. Yun, "An advanced approach for incremental flexible periodic pattern mining on time-series data," Expert Syst. Appl., vol. 230, no. 11, pp. 120697.1-120697.20, November 2023.

[8] T. S. M. Rawia and H. A. Rusli, "Toward developing a mining opinion model: conceptual framework," Acta Inform. Malaysia, vol. 2, no. 2, pp. 17-18, 2018.

[9] J. Pavlopoulos, A. Romell, J. Curman, O. Steinert, T. Lindgren, M. Borg, and K. Randl, "Automotive fault nowcasting with machine learning and natural language processing," Mach. Learn., vol. 113, no. 2, pp. 843-861, February 2024.

[10] M. Bartl, A. Mandal, S. Leavy, and S. Little, "Gender bias in natural language processing and computer vision: a comparative survey," ACM Comput. Surv., vol. 57, no. 6, pp. 139.1-139.36, February 2025.

[11] A. Joshi, R. Dabre, D. Kanojia, Z. Li, H. Zhan, G. Haffari, and D. Dippold, "Natural language processing for dialects of a language: a survey," ACM Comput. Surv., vol. 57, no. 6, pp. 149.1-149.37, February 2025.

[12] H. T. T. L. Pham and S. U. Han, "Natural language processing with multitask classification for semantic prediction of risk-handling actions in construction contracts," J. Comput. Civil Eng., vol. 37, no. 6, pp. 1.1-1.19, July 2023.

[13] S. Khan and M. Shaheen, "Wisdom mining: future of data mining," Recent Patents Eng., vol. 17, no. 1, pp. 2-11, January 2023.

[14] S. K. Pradhan, M. Jans, and M. Martin, "Getting the data in shape for your process mining analysis: an in-depth analysis of the pre-analysis stage," ACM Comput. Surv., vol. 57, no. 6, pp. 159.1-159.37, February 2025.

[15] O. Chambers, R. Cohen, M. R. Grossman, L. Hebert, and E. Awad, "Mining user study data to judge the merit of a model for supporting user-specific explanations of ai systems," Comput. Intell., vol. 40, no. 6, pp. 11.1-11.27, December 2024.

[16] H. Kim, M. Cho, H. Nam, Y. Baek, S. Park, D. Kim, B. Vo, and U. Yun, "Advanced incremental erasable pattern mining from the time-sensitive data stream," Knowl.-Based Syst., vol. 299, no. 9, pp. 1.1-1.18, September 2024.

[17] S. Kroeger, A. Rafles, P. Jordan, C. Soellner, and M. F. Zaeh, "Data model to enable multidimensional process mining for data farming based value stream planning in production networks," Prod. Eng., vol. 19, no. 2, pp. 307-327, April 2025.

[18] B. Bsir, N. Khoufi, and M. Zrigui, "Prediction of author's profile basing on fine-tuning bert model," Informatica, vol. 48, no. 1, pp. 69-78, January 2024.

[19] M. Tikhonova, V. Mikhailov, D. Pisarevskaya, V. Malykh, and T. Shavrina, "Ad astra or astray: Exploring linguistic knowledge of multilingual bert through nli task – corrigendum," Natural Lang. Eng., vol. 29, no. 3, pp. 554-583, April 2023.

[20] R. R. Sekar, T. D. Rajkumar, and K. R. Anne, "Deep fake detection using an optimal deep learning model with multi head attention-based feature extraction scheme," Visual Comput., vol. 41, no. 4, pp. 2783-2800, July 2024.

[21] W. Gao, H. Zhao, L. Li, and J. Zhu, "Text sentiment analysis based on ALBERT-HACNN-TUP model," Comput. Simul., vol 40, no. 05, pp. 491-496, July 2023.

Document Similarity Detection for Project Development Using Fused Interactive Attention Mechanisms

Chao Zhang[1], Ying Zhang[2], Gang Yang[3], Fan Hu[4]
State Grid Shanxi Electric Power Company, Taiyuan, 030021, China[1, 4]
State Grid Shanxi Electric Power Research Institute, Taiyuan, 030001, China[2, 3]

***Abstract*—This study introduces a novel multi-feature fusion model aimed at improving text similarity calculation in scientific and technological projects. The primary objective is to enhance the accuracy and efficiency of assessing text similarities, particularly in evaluating originality and identifying duplications in project submissions. To overcome the limitations of traditional text similarity methods (e.g., Vector Space Models, Latent Dirichlet Allocation, and TF-IDF) in capturing complex semantic and structural features, a hybrid model is proposed. The model combines word embeddings (word2vec and cw2vec), a Bi-LSTM network, and a multi-perspective convolutional neural network (MP-CNN) for effective feature extraction. Additionally, a fusion attention mechanism and interactive attention are incorporated to improve the extraction of semantic, contextual, and structural information. Experimental evaluation on two benchmark datasets demonstrates that the proposed model achieves an average precision of 0.75, a recall of 0.71, and an F1-score of 0.73, outperforming traditional methods (LDA, TF-IDF, Word2vec+Cosine) and deep learning baselines (Siamese-LSTM, MP-CNN) by more than 10% on average. These results confirm that the proposed architecture effectively balances semantic relevance and structural integrity, yielding superior similarity detection performance. The integration of advanced deep learning components—Bi-LSTM, MP-CNN, and attention mechanisms—substantially improves both the accuracy and efficiency of similarity evaluation, providing a more reliable and objective approach for scientific project assessment.**

Keywords*—*Text similarity; multi-feature fusion model; word2vec; cw2vec; MP-CNN; fusion attention mechanism; semantic extraction; project evaluation

I. Introduction

The duplication of scientific and technological projects is not only related to the smooth implementation of China's prospective research, but also has a profound impact on the orderly development of China's economy and culture. In order to promote the rational use of scientific research resources and funding, the scientific review of scientific and technological projects has become an important part of the scientific and technological program management system [1]. Therefore, in order to ensure the smooth development of original and innovative scientific research, project reviewers need to make more accurate judgments on the duplicity and similarity of the applied projects [2]. However, the duplicity review of scientific and technological projects is a complex process, with the year-on-year growth in the scale of scientific and technological project declaration, the project review methods used at this stage have been difficult to meet the current demand for originality review, and often need to be repeated by experts in various fields based on their own scientific research experience on the project and the related literature and patented technology screening, not only checking and checking for new accuracy fails to meet the requirements, but also seriously limits the efficiency of scientific and technological project review, resulting in the scientific and technological project review efficiency, which leads to a more accurate judgment of duplication and similarity. The review efficiency of science and technology projects has been seriously limited, which has hindered the management of science and technology programs.

Most of the traditional text similarity calculation methods apply literal repetition or probability models. For example, the traditional Vector Space Model1 (VSM) utilizes the theory of statistics to measure the similarity between texts based on the probability distribution of words. In [7], the authors utilize VSM to calculate the similarity between texts, adding keywords to avoid the removal of valid features. In recent years, a variety of methods have been proposed to enhance textual similarity detection. Traditional statistical approaches such as VSM, TF-IDF, and Latent Dirichlet Allocation (LDA) have laid the foundation for lexical and probabilistic modeling of text. However, these methods often fail to capture deep semantic and contextual information, leading to lower accuracy when dealing with complex or domain-specific documents. To overcome these limitations, deep learning-based approaches have been introduced. For instance, Siamese-LSTM models [3] learn sequence-level semantics through shared weight encoding, while CNN-based models such as MP-CNN [4] focus on extracting structural and local contextual features from paired sentences. Attention-based architectures like ABCNN [5] further refine feature alignment between text pairs, enabling improved sentence-level similarity detection. Despite these advances, existing models typically rely on single-level representations, either word-level or character-level, and rarely consider the integration of both semantic and structural perspectives. In contrast, our proposed model introduces a comprehensive multi-feature fusion strategy that simultaneously leverages word embeddings (word2vec and cw2vec), sequential encoding via Bi-LSTM, and multi-perspective convolution through MP-CNN, enhanced with fused and interactive attention mechanisms. This design enables the model to capture semantic relevance, structural integrity, and contextual dependencies in

an integrated framework, thereby addressing the key limitations of previous methods.

With the development of machine learning and deep learning technology, more and more researchers use deep learning technology to build models to study related tasks in the field of natural language processing. Vani Kl [6] et al. built a model for detecting plagiarism of academic ideas by taking plagiarism of ideas as a research object to address the increasing academic misconduct in the field of research and education. Velásquez [7] et al. proposed a plagiarism detection system called Document Copy Detector 3.0 (DOcument COpy Detector 3.0, DOCODE 3.0) to address the problem of academic plagiarism in educational institutions. Ehsan N et al. [8] addressed the problem of plagiarism being difficult to detect in cross-linguistic systems, and established a localized plagiarism detection model based on topic word retrieval and segment similarity assessment of hetero-linguistic sources as a research object. Arts S[9] et al. analyzed the limitations of the United States Patent Classification System (USPCS) in the detection of patent technology similarity, improved the text matching algorithm of the system, and proposed a text matching-based patent technology similarity detection algorithm. A text matching-based similarity analysis model for patented technologies. Sutoyo [10] et al. proposed a document plagiarism detection method based on a K-member grammar model with a screening algorithm and evaluated and selected the performance of the K-member grammar model's K-value with sliding window calculation. Scholar Choi S P M [11] et al. proposed an information retrieval-based text similarity detection algorithm that is capable of handling multilingual source documents and seamlessly integrates with existing learning management systems. The algorithm identifies potentially plagiarized phrases by employing information retrieval and sequence matching techniques, with parametric control to minimize false positives and negatives. Empirical evidence shows that the algorithm not only accurately and quickly identifies documents suspected of plagiarism, but also quantifies and visualizes the severity of plagiarism in data, thus providing scholars with a good aid in reviewing and assessing plagiarism.

Many researchers have achieved good results by not performing the pre-training task without text in order to take into account more comprehensive and underlying textual information. In [12], the authors does not carry out the pre-training of words, directly convolves the vectors represented by the unique hot code, fully mines and predicts the contextual information with the most original data, adds unsupervised region embedding, efficiently expresses the features of the text, and finally achieves better results on various tasks. Literature [13] utilizes convolutional neural networks to directly train and learn directly from the underlying characters of the text without text pre-processing, and applies the extracted features in various tasks of text processing, and achieves better results on large datasets. Literature [14] used fine-grained character-level text input to the convolutional neural network and then processed it using Long Short-Term Memory (LSTM), followed by applying the extracted features to a variety of languages to achieve better results, indicating that the model is able to obtain semantic information from the character-level input.

Secondly, to address the problem that entity relationships in the text of scientific and technological project declaration are difficult to be extracted effectively, an entity relationship extraction algorithm based on entity group co-occurrence rate is further proposed to realize high-quality entity relationship extraction in the text of scientific and technological project declaration; thirdly, to address the demand for similarity calculation of scientific and technological project declaration text and the structural characteristics of the text, a text matching model based on polytunnels is proposed to realize the comprehensive evaluation of semantic relevance of different components of the text. Finally, to address the problem of limited accuracy caused by only detecting the text as a whole or artificially setting the weights of each check item in the declaration text of scientific and technological projects, a semi-structured text similarity assessment method combining graph structural similarity and text matching degree is designed.

The remainder of this study is organized as follows: Section II presents the text preprocessing model and details the feature extraction process using word2vec and cw2vec embeddings. Section III introduces the proposed text similarity calculation model based on fused and interactive attention mechanisms. Section IV describes the experimental setup, datasets, and evaluation metrics, followed by a discussion of comparative results. Finally, Section V concludes the study and outlines directions for future work, including model optimization and potential generalization to other languages.

II. Text Preprocessing Model

Deep learning technology has achieved better results in various kinds of tasks, such as text similarity calculation, intelligent translation, sentiment classification, semantic analysis, etc., and has attracted many researchers at home and abroad. Deep learning utilizes multi-layer neural networks to extract deep features in the text, and in the field of natural language processing, it mainly uses convolutional neural networks and recurrent neural networks to extract text features, and due to the fact that the recurrent neural networks have memory units, the recurrent neural networks are more effective than the convolutional neural networks in various tasks in the field of natural language processing. As a whole, the text of a scientific and technological project declaration consists of structured document structure and semi-structured text data, and as a kind of text data with special text features, this text type has a big difference from free text data in terms of syntax, wording, and article organization, thus presenting obvious text structural features and semantic features.

Word vectors are the prerequisite for the calculation of semantic similarity of text, so it is necessary to pre-process and pre-train the text to obtain high-quality word vectors, the process is shown in Fig. 1. Preprocessing is to segment the text in the corpus and training set and remove the stop words in the text, and pre-training is to convert the word sequences into feature vectors that can be recognized by the computer.

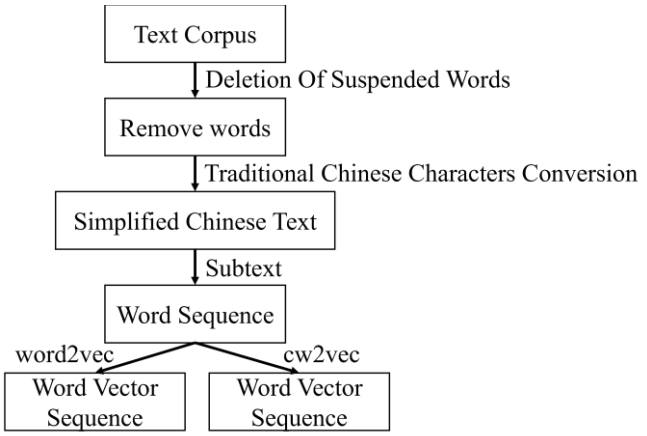


Fig. 1. Text preprocessing model.

A. Text Segmentation

Chinese word segmentation is the process of splitting a Chinese text into words according to semantic criteria and combining the results into a new sequence. Since English words are separated by spaces, each word can be used as a semantic unit, so English text segmentation is less difficult. On the other hand, Chinese text consists of consecutive Chinese characters connected together, and there is no separating mark between each word, so the segmentation of Chinese text is a basic and important step, and the accuracy of this process has a large impact on the subsequent related tasks and an accurate and fast segmentation algorithm is needed before the model training. On the issue of selecting a text segmentation model, the word labeling-based conditional random field model is the most used segmentation model, which is a model that uses word construction rules, and has a higher recall rate for unregistered words, but at the same time, it also generates more segmentation errors. Segmentation models using word annotations also require the addition of a complex denoising process in subsequent tasks. For the binary syntactic participle model, it only recalls the words present in the word list, and combined with the new word discovery algorithm, it can effectively alleviate the recall problem of high-frequency unregistered words. Assume that the sentence $\boldsymbol{T = w_1 w_2 w_3 \ldots w_n}$ has completed the disambiguation operation, where w is the n words composing sentence T, sentence T is changed to the original sentence $\boldsymbol{S = c_1 c_2 c_3 \ldots c_n}$, after passing through the noise channel without disambiguation, where $\boldsymbol{c}$ is the Chinese character in the sentence. Calculation using the Bayesian formula gives in Eq. (1):

$$\boldsymbol{P(T \mid S) = \frac{P(T)P(S|T)}{P(S)} \propto P(T) = P(w_1 w_2 w_3 \ldots w_n)}$$

$$= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2)\ldots P(w_n \mid w_1 w_2 \ldots w_{n-1}) \quad (1)$$

Assume that the current word is only related to the previous work:

$$P(T) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_2)\ldots P(w_n \mid w_{n-1}) \quad (2)$$

This is the binary grammatical disambiguation model [Eq. (2)]. Add the interpolation smoothing calculation, as shown in Eq. (3):

$$P(w_n \mid w_{n-1}) = \alpha P(w_n \mid w_{n-1}) + (1-a)P(w_n) \quad (3)$$

where, $\boldsymbol{\alpha}$ is taken as 0.7. Jieba is one of the most active Chinese word splitter tools in China with a large number of users, which provides various functions such as word splitting, keyword, extraction, word labeling and so on. The participle function mainly has three modes: search engine mode, full mode, and exact mode, which can be applied in search engine, text, sentence and other scenarios that need participle. And the exact model is most suitable for use in the task of text analysis, therefore, the experiments of the text selected jieba participle tool to assist in participle.

B. Word-CW2VEC Model

Representing text using computer-recognizable word vectors can effectively alleviate the situation where the text is not trainable due to its variable length, large data dimension, and complex structure. Utilizing trainable word vectors can preserve similar features at the semantic level, mapping high-dimensional text data to low dimensions while avoiding dimensional catastrophe. Bengio et al. [15] used a neural network language model to train the data, learning the feature representations of the words as word vectors through the hidden layer, which learns the semantic information through the neural network.

Word2vec is a Google open-source tool for calculating word vectors. The two models CBOW and Skip-Gram, used in the tool for calculating the generated word vectors are proposed by Mikolov et al. [24]. The models simplify the Neural Network Language Model (NNLM) and design an accelerated training strategy to allow the model to be efficiently trained on massive training sets. The word vectors trained on a large amount of text data are able to represent the semantic relationships between words and tap into deep features. Because of their good performance and performance, the two models CBOW and Skip-Gram are widely used and have achieved good results on many natural language processing tasks. CBOW's core concept is to use words within a specified distance around a central word as context to model and predict the likelihood of the central word using a linear model. The architecture is shown in Fig. 2. The CBOW model is improved from NNLM by abandoning the strategy of a nonlinear hidden layer and vector splicing, which affects the training efficiency of NNLM, and mapping the word vectors to the same location.

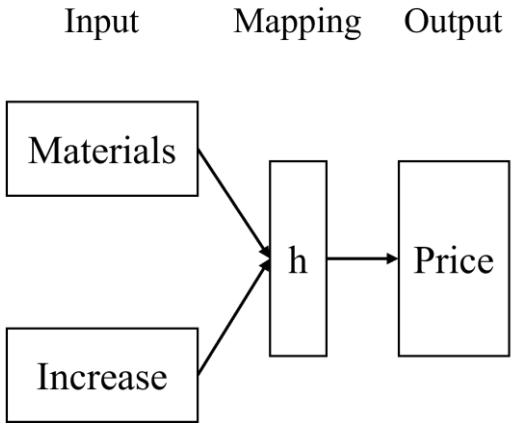


Fig. 2. CBOW model.

For a text matrix $W = (w_1, w_2, \ldots, w_n)$, CBOW utilizes the mapping layer e to sum the word vectors in the context c expressed as Eq. (4):

$$\boldsymbol{h} = \frac{\mathbf{1}}{\boldsymbol{n-1}} \sum_{\boldsymbol{w_i \in c}} \boldsymbol{e(w_i)} \tag{4}$$

The center word is then predicted, and the weights are constantly updated by maximizing the conditional probability with the contextual representation h, maximizing the conditional probability as:

$$\boldsymbol{L} = \sum_{\boldsymbol{(w,c)\in D}} \boldsymbol{log\, P(w \mid c)} \tag{5}$$

$$\boldsymbol{P(w \mid c)} = \frac{\boldsymbol{exp(e'(w)^T h)}}{\sum_{\boldsymbol{w' \in V}} \boldsymbol{exp\,(e'(w')^T h)}} \tag{6}$$

where, $\boldsymbol{e}$ is the mapping function, w is the center word, and $\boldsymbol{c}$ is the contextual information [Eq. (5) and Eq. (6)].

CW2VEC is a method proposed by Cao [16] and others to decompose Chinese text strokes in order to extract deeper and finer-grained information, and utilize Chinese strokes for model training and feature representation. The n-gram language model is utilized to mine the associations and semantic information on text morphology, and the sliding window is changed by adjusting the size of n to extract the semantic information of different granularity of strokes. The processing flow of the n-gram model based on strokes is shown in Fig. 3.

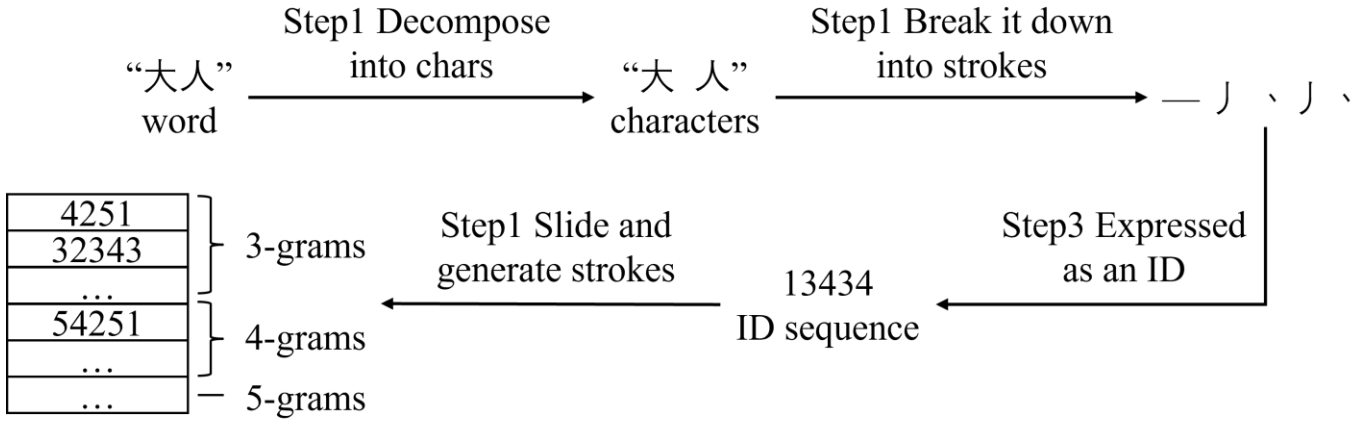


Fig. 3. Stroke-based n-gram modeling.

In Fig. 4, the model structure of cw2vec is shown; the basic idea is similar to Skip-Gram, both of them use the center word to predict the context information. The difference is that cw2vec uses the n-gram language model of strokes to represent the context information.

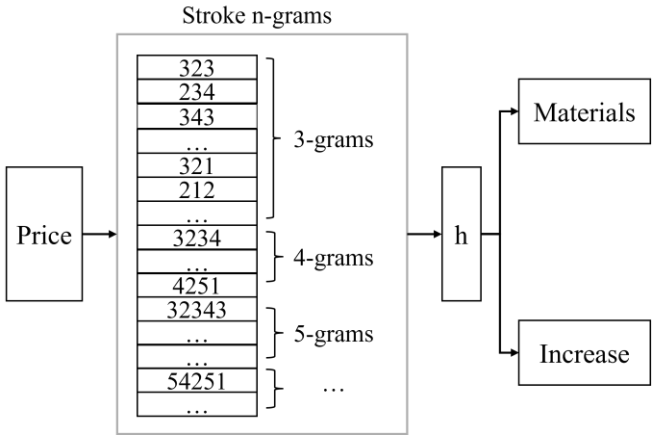


Fig. 4. CW2VEC model structure.

The context vector of the center word w is the sum of the feature vectors derived from the stroke n-gram model, computed as Eq. (7):

$$\boldsymbol{h} = \sum_{\boldsymbol{q \in S(w)}} \boldsymbol{e(q)} \tag{7}$$

where, S(w) is the set of stroke features of the center word computed using the n-gram model. The model is trained using the conditional probability of maximizing the center word with the words in the context, with the following Eq. (8) and Eq. (9):

$$\boldsymbol{P(w \mid w_j)} = \frac{\mathbf{exp}(\boldsymbol{e'(w)^T e(w_j)})}{\sum_{\boldsymbol{w/wV}} \mathbf{exp}\,(\boldsymbol{e'(w')^T e(w_j)})} \tag{8}$$

$$\boldsymbol{P(w \mid h_i)} = \frac{\mathbf{exp}(\boldsymbol{e'(w)^T h_i})}{\sum_{\boldsymbol{w' \odot V}} \mathbf{exp}\,(\boldsymbol{e'(w')^T h_i})} \tag{9}$$

For deep learning models, inputs are very important, and although it is possible to design structures with high performance, it is difficult for the model to work well if the input information is limited. In the experiments, it was found that word-level-based results are average, which is because the quality of the participle has a great impact on the model. To explore deeper semantic and structural information in the text, this study introduces a word embedding method combining word2vec and cw2vec, integrating three input sources from stroke sequences, word sequences, and word sequences, where the word2vec word embedding is inputted into the model while the cw2vec word embedding based on strokes is used in the other input channel, and then the two results are fused. This method effectively alleviates the problem of poor quality of word separation in the model. To summarize, the text structure feature extraction structure of this study is shown in Fig. 5.

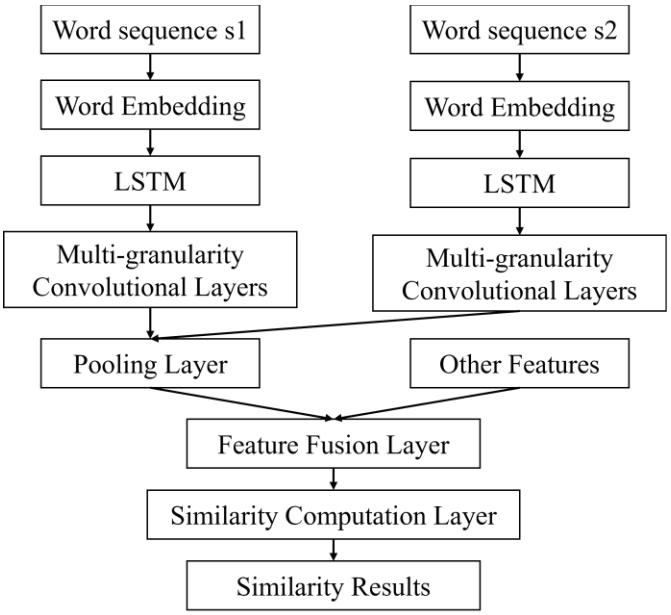


Fig. 5. Text structure feature extraction module.

III. Text Similarity Calculation Model Based on Fused Attention Mechanism

The similarity calculation of text is generally divided into three steps: first of all, the corpus should be preprocessed, including de-duplication, text segmentation, pre-training, and other steps. Then the text representation model is used to extract the feature representation containing semantics in the text, and finally, the similarity is obtained by training the model according to the extracted features.

A. Text Structure Feature Extraction Module Design

In this study, the proposed multi-feature fusion model for text similarity calculation of science and technology projects draws on the idea of Siamese structure, adopts word2vec and cw2vec word embeddings with different granularity as inputs, and jointly extracts the semantic information of the text with a Bi-LSTM network; effectively extracts the structural and word order information of the text through multi-granularity convolution and corresponding pooling; proposes the LSF feature The computation method of LSF features is proposed and proved to be effective. With the advantages of the two models, the extracted features are effectively fused, the semantic features are better preserved, and the deep-level features are mined. Finally, the text similarity calculation module for science and technology projects is designed, and the details of parameter settings are explained.

The whole model is divided into a word embedding module, a text structure feature extraction module (CNN model), a text semantic information extraction module (Bi-LSTM model), an attention mechanism module, a feature fusion module, and a similarity computation module. The structure of the similarity computation model with CNN fusion attention mechanism is shown in Fig. 6.

In this study, we draw on the MPCNN [17] (Multi-Perspective Convolutional Neural Networks) model to design the network structure, and use the convolution of multi-granularity to explore more deep features hidden in the present.

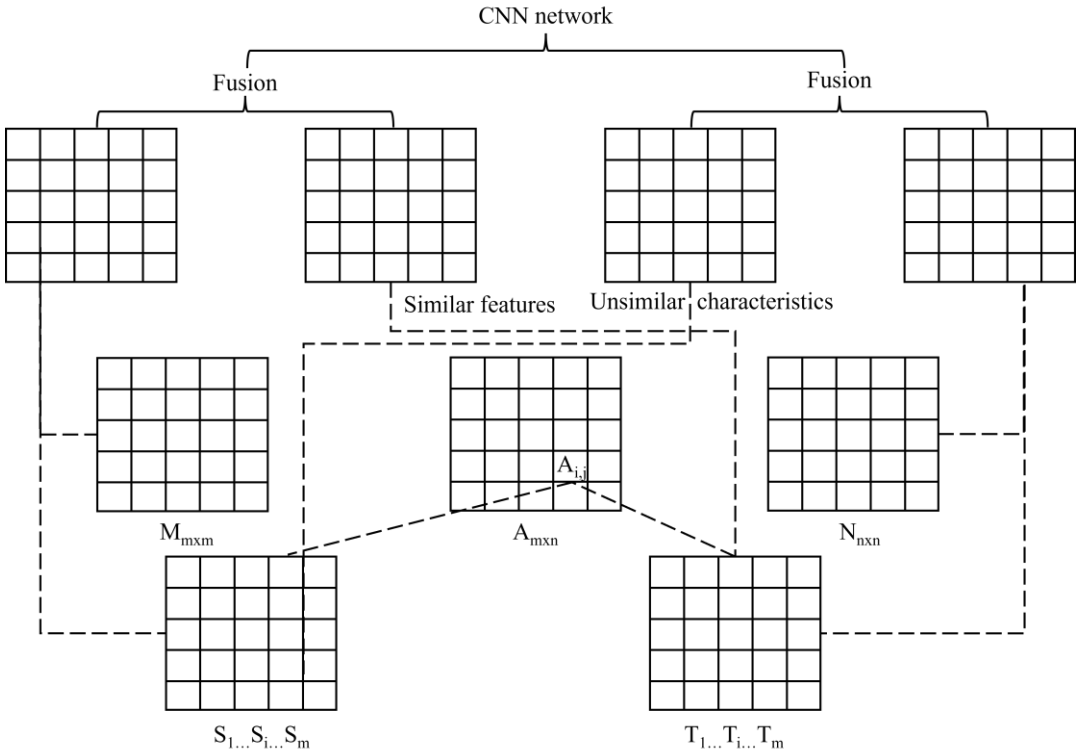


Fig. 6. CNN fusion attention mechanism model.

The structure of CNN-based model for calculating text similarity of science and technology projects is shown in Fig. 6. After the input text data is preprocessed with preprocessing operations such as word splitting and deactivation, the text is trained into word vectors and combined into a text matrix using the word2vec tool. Then the LSTM network is used to further train the word vector representation so that the word vectors contain more semantic information.

This study, inspired by Shen T, proposes an effective fusion of semantic features, contextual structural interaction features, and LSF features extracted from text through the text semantic, structural, and LSF feature extraction modules [18], and the fusion gate feature fusion method is designed, with the following computational equations [Eq. (10) to Eq. (12)]:

$$\boldsymbol{z_t} = \boldsymbol{sigmoid}(\boldsymbol{W_z} \cdot [\boldsymbol{h_{t-1}}; \boldsymbol{x'_t}] + \boldsymbol{b_z}) \quad (10)$$

$$\boldsymbol{r_t} = \boldsymbol{sigmoid}(\boldsymbol{W_r} \cdot [\boldsymbol{h_{t-1}}; \boldsymbol{x'_t}] + \boldsymbol{b_r}) \quad (11)$$

$$\boldsymbol{h_t} = (\boldsymbol{1} - \boldsymbol{z_t}) \times \boldsymbol{h_{t-1}} + \boldsymbol{z_t} \times \tilde{\boldsymbol{h}}_{\boldsymbol{t}} \quad (12)$$

The method fuses $\boldsymbol{h_{t-1}}$ and $\boldsymbol{x_t}$ into $\boldsymbol{h_t}$., where $\boldsymbol{z_t}$ is used to measure the degree of feature fusion, $\boldsymbol{x'_t}$ is the vector obtained by projection transformation of $\boldsymbol{x_t}$, W is the trainable parameter matrix, b is the trainable parameter vector, and the [;] symbols are used to splice the two vectors. Fusion gate method can accelerate the flow of information efficiently, and further integrates the computation results with a layer of multilayer. The fusion gate method can effectively accelerate the information flow, further integrate the computational results with a multilayer perceptual machine, and finally get the vectors with the same dimensions as the features before fusion.

B. Integration of Interactive Attention Mechanisms

In this study, drawing on the ideas of the ABCNN model and the LDC model [19], we introduce the interaction attention mechanism into the CNN model, and propose an improved scheme to construct the interaction attention matrix and take into account the similarities and dissimilarities through the orthogonal decomposition of the matrix. The model structure is shown in Fig. 7:

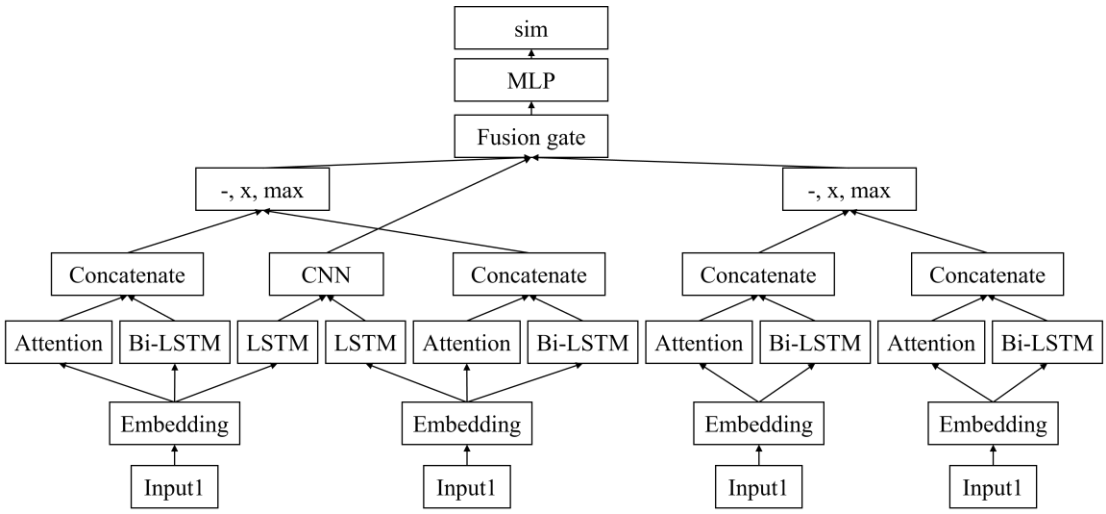


Fig. 7. Multi-feature fusion similarity calculation model.

For text pairs S and T, the text lengths are set as m and n, respectively, in order to obtain the interactive attention vector $\hat{s}_i$ (which implies similar and dissimilar components) of the word vector s_i in the text s, it is necessary to compute the similarity matrix $A_{m\times n}$, which is calculated as Eq. (13):

$$a_{i,j} = \frac{{s_i}^T t_j}{\|s_i\| \cdot \|t_j\|} \quad \forall s_i \in S, t_j \in T \quad (13)$$

where, $a_i, j \in A_{m\times n}$,is an element in $A_{m\times n}$, $s_i (i < m)$ is a word vector in text S, and $t_j (j < n)$ is a word vector in text T. The interactive attention vector of text s can be computed from the text representation matrix T and the similarity matrix $A_{m\times n}$ with the following Eq. (14):

$$\hat{s}_i = f_{match}(s_i, T) = \frac{\sum_{j=k-w}^{k+w} a_{i,j} t_j}{\sum_{j=k-w}^{k+w} a_{i,j}} \quad (14)$$

where, $k = \text{argmax} x_j a_{i,j}$. $\hat{s}_i = f_{match}(s_i, T)$, $\hat{s}_i$ is the ith vector in the interaction attention matrix of the text matrix S. w is a variable parameter, which is represented by w values near the maximum of the similarity matrix, i.e., the local interaction semantics.

Each vector is orthogonally decomposed into two parts of the geometric space parallel and perpendicular, where parallel is the similar part and perpendicular is the similar part, and where parallel is the similar part and perpendicular is the dissimilar part. The decomposition equation is Eq. (15) and Eq. (16):

$$s_i^+ = \frac{s_i \cdot \hat{s}_i}{\hat{s}_i \cdot \hat{s}_i} \hat{s}_i \quad parallel \quad (15)$$

$$s_i^- = s_i - s_i^+ \quad perpendicular \quad (16)$$

Using the orthogonal decomposition described above, the interactive attention matrix of the text is decomposed into similar and dissimilar matrices, denoted as $S^+ = [s_1^+, \ldots, s_i^+ \ldots, s_m^+]$ and $S^- = [s_1^-, \ldots, s_i^- \ldots, s_m^-]$. Both the dissimilar and similar parts are strongly related to each other, and it is difficult to determine the degree of similarity between similarly shaped but meaningfully different texts if only the similar parts are considered. When the similar and different parts

are considered at the same time, it can be historically good to determine the similarity gui between such texts. Therefore, the model in this study synthesizes the similar component matrix and the different component matrix into one feature vector $\vec{S}$ and $\vec{T}$, and the synthesis function is [Eq. (17) and Eq. (18)]:

$$\vec{S} = f_{comp}(S^+, S^-) \quad (17)$$

$$\vec{T} = f_{comp}(T^+, T^-) \quad (18)$$

where, f_{comp} is a combinatorial function. For the convolution operation, a list of convolution kernels w_o is defined, and each convolution kernel has the shape of $d \times h$, where d is the dimension of the word vector and h is the window size. In Eq. (17) and Eq. (18), each convolution kernel is applied to two channels from similar and dissimilar to generate a feature. The process is shown in the following Eq. (19):

$$c_{o,i} = f\left(w_o * S^+_{[i:i+h]} + w_o * S^-_{[i:i+h]} + b_o\right) \quad (19)$$

where, the $A * B$ operation adds all the elements in B with the corresponding weights in A, $S^+_{[ii+h]}$ and $S^-_{[ii+h]}$ denote the parts from S^+ and S^-, b_o is a bias term, and f is a nonlinear function. Finally, the similar and dissimilar features extracted by the CNN are spliced as extracted features and fused with other features to compute the similarity after the similarity computation layer.

IV. Results and Discussion

A. Experimental Data Collection

The experiments in this study are implemented using the open-source machine learning framework keras, which is rich in documentation, easy to use, and simple to model, and has attracted a large number of developers. In the experiment, an NVIDIA GTX 1050Ti is used as an auxiliary tool for model training, and through the Unified Computing Architecture technology introduced by NVIDA, the graphics card can be used to accelerate the training speed by utilizing APIs on GPUs (General-Purpose GPUs). The jieba segmentation tool is used to segment Chinese words, the Skip-Gram model of the word2vec tool is used to train word vectors, and cw2vec is used to train fine-grained word vectors.

Since there is no publicly available R&D dataset for science and technology projects, two datasets are chosen in this study to validate the experiments. Dataset 1 is the dataset of the Financial Intelligence NLP service tournament of Ant Financial Competition. The dataset is given 100,000 pairs of labeled data, all of which are two paragraphs of customer service and user Q&A, which contain synonymous pairs and non-synonymous pairs, and the algorithm is used to determine whether the same semantics are represented. Dataset 2 is the ChineseSTS similarity training set organized by Xi'an University of Science and Technology (XUST), in which 27,000 pairs of texts are classified into six similarity levels, in which the similarity of completely similar pairs is 5, and the similarity of dissimilar pairs is 0. However, the distribution of the pairs of texts in the training set is not balanced, and 90% of the pairs of texts have similarity of 0 or 5, which makes no sense for the training of the model. Therefore, 6000 text pairs are selected for model training and model performance evaluation. In order to harmonize with dataset 1, the similarity of text pairs in dataset 2 with similarity greater than 2 is set to 1, and the similarity of text pairs with similarity less than or equal to 2 is set to 0, which is used for the training and testing of the model.

B. Evaluation Indicators

In the study of text similarity calculation models for science and technology projects, it is very necessary to evaluate the model, and the evaluation index can measure the goodness of a model. In order to facilitate the evaluation of the performance of the model, this study evaluates the science and technology project text similarity calculation as a binary classification problem, which maps the model prediction results to the set {0,1}, with 0 representing that the two texts are not similar and 1 representing that the two science and technology project texts are similar. The commonly used evaluation metrics for binary classification problems are Precision, Accuracy, Recall, and F1-Measure. The following classes are defined according to the categories predicted by the model and the real categories of the samples:

1) TP (True Positive): model predicts a positive class and the sample is truly labeled as a positive class.

2) TN (True Negative): model predicts a negative category and the sample is truly labeled as a negative category.

3) FP (False Positive): model predicts a positive class, the real labeled as a negative sample.

4) FN (False Negative): model predicts a negative class, the real labeled as a positive class of samples.

where, the total number of samples is the sum of the four classes TP, TN, FP and FN. As shown in Table I, the confusion matrix can be used to describe the relationship between TP, TN, FP, and FN.

TABLE I. Confusion Matrix

	Similarity, positive	**Unsimilar, negative**
Physical resemblance	TP, Positive predicted to be positive	FP, Negative predicted to be positive
Actual dissimilarity	FN, Positive predicted to be negative	TN, Negative predicted to be negative

Accuracy and recall are widely used in the field of statistical classification and can be used to evaluate the accuracy and comprehensiveness of each prediction of a model. Among them, the accuracy rate is the ratio of the number of correct predictions made by the model to the total number of samples, and the equation is as follows [Eq. (20)]:

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad (20)$$

Accuracy evaluates the goodness of the model from the perspective of the model's prediction results, and is the proportion of the samples predicted to be positive classes that are truly labeled as positive classes, which measures the model's checking accuracy, with the following Eq. (21):

$$\text{Precision} = \frac{TP}{TP+FP} \quad (21)$$

Recall analyzes the goodness of the model from the point of view of the labeling results (see Eq. 21), describes the ratio of the number of samples predicted to be positive by the model to the number of samples labeled to be positive, and measures the model's checking rate, the equation is as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{22}$$

In general, recall and precision are negatively correlated (Eq. 22), and the F1 value, which is an indicator of the overall evaluation of the model, is the harmonic mean of recall and precision, with the following Eq. (23):

$$\text{F1} = \frac{2^{*}\text{Precision}^{*}\text{Recall}}{\text{Precision+Recall}} \tag{23}$$

C. Test Results

In order to verify that the model proposed in this study has good results, several models commonly used at present are implemented as comparison models, including traditional similarity calculation models and deep learning-based similarity calculation models.

LDA: The traditional LDA algorithm [20] is used to calculate the similarity of texts.

TF-IDF: The four text keywords in the two texts are extracted separately, and the cosine value is calculated as the similarity of the texts using the word embedding representation vectors of the keywords [21];

word2vec-cos: the representation vectors of words are trained using word2vec and combined with the cosine value to calculate the text similarity;

Siamese-LSTM: used to verify the validity of textual semantic features, the input is a text matrix of word vectors trained with the word2vec framework [22];

Siamese-CNN: based on the feature extraction model proposed by Kim, based on the Siamese framework and CNN network, with the input being a matrix representation of the sentence itself and convolved in full dimension;

MP-CNN: used to verify the validity of text structure features, the input is text matrix, and the text similarity is computed after extracting features with different granularity of convolution kernels and multiple pooling methods [23];

Siamese-LSTM-cw2vec: used to validate the effectiveness of cw2vec word embedding, the input of the model is the text matrix with stroke granularity trained with the cw2vec framework;

Our Model: the algorithm proposed in this study, the semantic, contextual structure of the text extracted from the model, the interaction between the text features using fusion gate fusion through the similarity calculation layer to calculate the text similarity.

In order to investigate the performance difference between the text similarity computation model designed and implemented in this study and the mainstream model, and to verify the advantage of multi-feature fusion over single feature, seven sets of comparison tests are set.

The average percentage error MSPE is shown in Fig. 8.

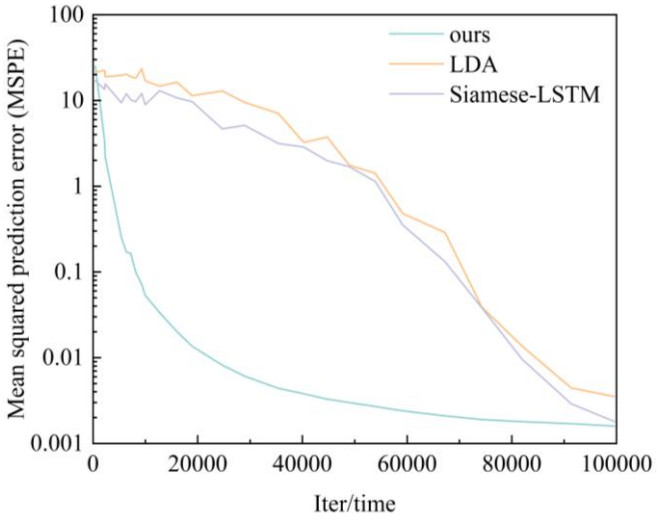


Fig. 8. Model percentage error plot.

TABLE II. COMPARISON OF RESULTS OF DIFFERENT MODELS IN DATASET 1

Model	Precision	Recall	F1
LDA	0.470	0.487	0.478
TF-IDF	0.433	0.468	0.450
Word2vec+cos	0.503	0.524	0.513
Siamese-LSTM	0.554	0.566	0.560
MP-CNN	0.563	0.582	0.572
Siamese-LSTM-cw2vec	0.536	0.572	0.553
Our Model	0.649	0.624	0.636

As can be seen from Table II and Table III, the F1 value of the model proposed for text is improved by about 0.2 compared to the traditional LDA model and the TF-IDF model, and also improved by about 0.1 compared to the Word2vec-cos model. In terms of the text feature representation, pre-trained word embeddings improve the model's ability to represent word features, and the word features trained by the Bi-LSTM network contain higher-order contextual semantic information than the n-gram model with higher order contextual semantic information. To further illustrate the model's computational efficiency, Fig. 9 presents a comparative analysis of time cost across different models. It can be observed that, despite the multi-component architecture, our approach maintains a competitive processing time, demonstrating that the performance gains do not come at the expense of excessive computational overhead. The similarity computation method based on the fusion attention mechanism proposed in this study improves the performance over the existing mainstream methods, which is mainly due to the model's effective fusion of the semantic information of the text and the structural information of the context. By effectively fusing multiple features extracted from the model using the fusion gate to retain as much information as possible from all the features, the performance is better than that of a single feature.

TABLE III. COMPARISON OF RESULTS OF DIFFERENT MODELS IN DATASET 2

Model	Precision	Recall	F1
LDA	0.551	0.568	0.559
TF-IDF	0.579	0.595	0.587
Word2vec-cos	0.594	0.612	0.603
Siamese-LSTM	0.652	0.672	0.662
MP-CNN	0.663	0.657	0.660
Siamese-LSTM-cw2vec	0.682	0.679	0.681
Our Model	0.749	0.714	0.731

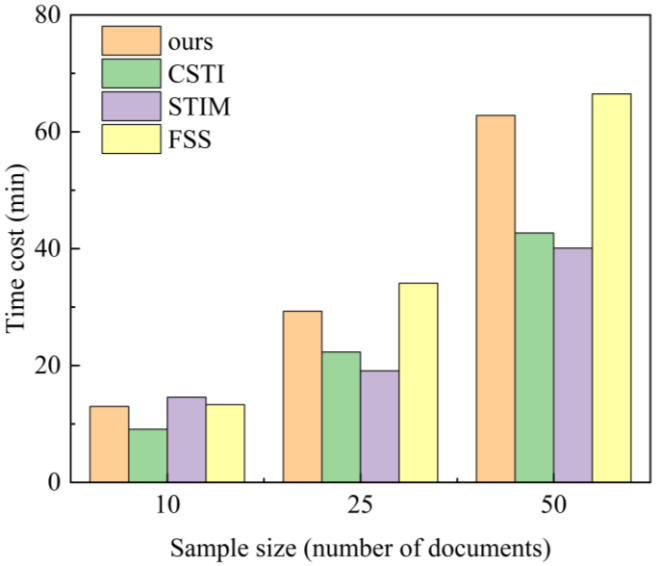


Fig. 9. Comparison chart of the time cost of our method.

Siamese-LSTM model applies the LSTM network with shared weights to the input coding of two texts under the Siamese framework, and the model is easy to train, which is a commonly used method in the field of text similarity computation with better results. However, some of the interaction features are missing, and the introduction of the product of feature vectors, variance, cosine value and Manhattan distance to amplify the dissimilarities of the texts can effectively alleviate the lack of interaction features. The input of the Siamese-LSTM-cw2vec model is the embedding of the stroke granularity of cw2vec, which has a better effect, proves the effectiveness of fine-grained word embedding, and explains that this study proposes the model incorporating cw2vec word embeddings. The MP-CNN model as a single-feature comparison model extracted some features. Word2vec-cos method is less effective because the simple representation of the text as a feature vector with word2vec and then calculate the similarity with the cosine value cannot fully take into account the complex semantic and syntactic information of the Chinese text. LDA is a topic model, which is essentially a bag-of-words model-based LDA is a topic model, which is essentially a model based on bag-of-words model to deal with long text, while the training set constructed in this study has shorter text, mostly belonging to the same type of topics, which can't take advantage of the performance in the experiment, and the performance is not good. The performance of TF-IDF method in the training set is not good because measuring the importance of a word simply by its "word frequency" is not comprehensive enough, and words that affect the semantics may appear more often because of their importance. The reason is that simply measuring the importance of a word by its "word frequency" is not comprehensive enough, and words that affect the semantics of a word because of their importance may appear less frequently, which does not reflect the positional information of the word and does not involve semantic features. It also relies heavily on the corpus, and needs to select a large number of high-quality corpora that match the training task.

V. CONCLUSION

The present study proposes a multi-feature fusion model for text similarity calculation in scientific and technological projects, effectively integrating semantic, contextual, and structural features to enhance accuracy. Compared with traditional similarity calculation methods, the proposed model achieves significant performance improvements, demonstrating its effectiveness in preserving both semantic meaning and contextual relationships. The combination of word2vec and cw2vec embeddings, together with Bi-LSTM and multi-perspective convolutional neural networks (MP-CNN), enables comprehensive feature extraction and fusion at multiple linguistic levels. Furthermore, the incorporation of fusion and interactive attention mechanisms enhances the model's capability to capture both shared and distinctive patterns across texts, thereby improving overall similarity detection performance.

Experimental results on two benchmark datasets indicate that the proposed model consistently outperforms traditional approaches such as LDA and TF-IDF, as well as advanced deep learning models including Siamese-LSTM and MP-CNN. An average F1-score improvement exceeding 10% over baseline methods highlights the model's ability to capture nuanced semantic dependencies and structural correlations.

Nevertheless, the multi-component architecture introduces additional computational complexity. To address this challenge, future work should focus on optimizing parameter efficiency through lightweight attention mechanisms, dimensionality reduction strategies, and GPU-based parallelization techniques, thereby improving the feasibility of real-time and large-scale deployment. In addition, it is valuable to investigate the model's generalization across multiple languages and domains. Given that cw2vec is primarily designed for Chinese text, subsequent research may incorporate multilingual embeddings (e.g., mBERT or cross-lingual transfer learning methods) to extend applicability to other linguistic contexts and scientific disciplines.

In summary, the proposed multi-feature fusion framework effectively balances semantic relevance, structural integrity, and contextual comprehension. With continued optimization and multilingual adaptation, the model has the potential to evolve into a scalable, domain-independent solution for intelligent and real-time text similarity assessment.

REFERENCES

[1] Chandrasekaran, Dhivya, and Vijay Mago. "Evolution of semantic similarity—a survey." ACM Computing Surveys (CSUR) 54.2 (2021): 1-37.

[2] Pearce, Joshua M. "Economic savings for scientific free and open source technology: A review." HardwareX 8 (2020).

[3] Shih, Chin-Hong, et al. "Investigating siamese lstm networks for text categorization." 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2017.

[4] Ding, Peng, et al. "A novel discrimination structure for assessing text semantic similarity." Journal of Internet Technology 23.4 (2022): 709-717.

[5] Yin, Wenpeng, et al. "Abcnn: Attention-based convolutional neural network for modeling sentence pairs." Transactions of the Association for computational linguistics 4 (2016): 259-272. Vani K, Gupta D, Detection of Idea Plagiarism Using Syntax-semantic ConceptExtractions with Genetic Algorithm[J]. Expert Systems with Applications, 2017, 73 :11-26.

[6] Vel¡squez JD, Covacevich Y, Molina F,et al. DOCODE 3.0 (DOcument COpyDEtector): A System for Plagiarism Detection by Applying an Information FusionProcess from Multiple Documental Data Sources[J. Information Fusion, 2016, 27:64-75.

[7] Ehsan N, Shakery A. Candidate Document Retrieval for Cross-lingual PlagiarismDetection Using Two-level Proximity Information[J]. Information Processing &Management,2016,52(6): 1004-1017.

[8] Arts S, Cassiman B, Gomez JC. Text Matching to Measure Patent Similarity[J]Strategic Management Journal,2018,39(1):62-84.

[9] Sutoyo R, Ramadhani I, Ardiatma A D, et al. Detecting Documents Plagiarism UsingWinnowing Algorithm and K-gram Method{C}//2017 EEE International Conferenceon Cybernetics and Computational Intelligence (CyberneticsCom). Phuket: IEEE,2017:67-72.

[10] Choi S P M, Lam S S. IChecker: An Efcient Plagiarism Detection Tool for LearningManagement SystemsJ. International Journal of Systems and Service-OrientedEngineering(IJSSOE),2016,6(3): 16-31.

[11] Johnson R, Zhang T. Semi-supervised convolutional neural networks for text categorization via region embedding[C].Proceedings of the 28th International Conference on Neural Information Processing Systems,2015:919-927

[12] Kim Y, Jernite Y, Sontag D, et al. Character-Aware Neural Language Models[C].The 30th AAAI Conference on Artificial Intelligence,2016:2741-2749

[13] Turian J, Ratinov L, Bengio Y. Word representations:A simple and general method for semi-supervised learning[C].Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,2010:384-394.

[14] Bengio Y, Ducharme R, Vincent P,et al.A neural probabilistic language model[J]. Journal of machine learning research,2003,3(2):1137-1155.

[15] Cao S, Lu W, Zhou J, et al.cw2vec:Learning chinese word embeddings with stroke n-gram information[C].The 32th AAAI Conference on Artificial Intelligence, 2018:5053-5061.

[16] He H, Gimpel K, Lin J. Multi-perspective sentence similarity modeling with convolutional neural networks[C].Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015:1576-1586.

[17] Shen T, Zhou T, Long G, et al. Disan: Directional self-attention network for rnn/cnn-free language understanding[C].The 32th AAAI Conference on Artificial Intelligence.2018

[18] Wang Z, Mi H, Ittycheriah A. Sentence similarity learning by lexical decomposition and composition[C].International Conference on Computational Linguistics, 2016:1340-1349

[19] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J].Journal of machine Learning research,2003,3(1):993-1022

[20] Grootendorst, Maarten. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." arXiv preprint arXiv:2203.05794 (2022).

[21] Turian J, Ratinov L, Bengio Y. Word representations: A simple and general method for semi-supervised learning[C].Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics,2010:384-394

[22] Yin W, Schütze H, Xiang B, et al. Abcnn: Attention-based convolutional neural network for modeling sentence pairs[J].Transactions of the Association for Computational Linguistics,2016,4(1):259-272.

[23] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems 26 (2013).

A Comparison Between FAHP-TOPSIS and FAHP-FTOPSIS Methods for Selecting the Best Products for Home-Based Sellers: A Performance Analysis

Selvia Lorena Br Ginting[1]*, Zulaiha Ali Othman[2]
Center for Artificial Intelligence Technology (CAIT)-Faculty of Information Science and Technology,
University Kebangsaan Malaysia, Bangi 43600, Malaysia[1, 2]
Computer System Study Program-Faculty of Engineering and Computer Science,
Universitas Komputer Indonesia, Bandung, Indonesia[1]

***Abstract*—This study discusses common problems faced by home-based sellers in determining the right product ideas to sell. To overcome this problem, a method is needed that can help home-based sellers to choose the right product. Therefore, a decision support system using a multi-criteria decision-making technique with a hybrid approach was applied, which integrates the FAHP-TOPSIS and FAHP-FTOPSIS methods in the product selection process. The analysis results show that the FAHP-TOPSIS method is more effective in producing product rankings, with alternative A5948 ranking first with a score of 0.946. Meanwhile, the FAHP-FTOPSIS method also placed the same alternative in first place with a score of 0.679. The findings in the ranking analysis showed that the addition of fuzzy did not affect the rankings but did affect the score value of the alternatives. Sensitivity Analysis using Mean Absolute Deviation (MAD), Mean Square Error (MSE), and Spearman Correlation (SC) was conducted. FAHP-TOPSIS performed best at Weight 1 (MAD 89, MSE 18.486, SC 0.972) and excelled at Weight 3 (MAD 144, MSE 51.791, SC 0.997), though more volatile at other weights. Overall, at the base weight (Weight 1), TOPSIS shows the best ranking stability (low MAD/MSE, high SC), while with shifted weights (especially Weight 3), FTOPSIS better maintains ordering (SC ≈ 1) despite higher error at Weight 2. Practically, TOPSIS suits baseline scenarios; FTOPSIS is more robust under weight variations, with error variance control still necessary. These findings provide a practical guideline: use FAHP-FTOPSIS when preferences are uncertain, and use FAHP-TOPSIS when preferences are clear. The resulting rankings can be directly adopted by sellers to prioritize and select products with confidence.**

***Keywords*—Home-based sellers; product selection; FAHP-TOPSIS; FAHP-FTOPSIS; sensitivity analysis**

I. Introduction

The rapid growth in internet availability and improvements in the quality of high-speed internet services have created significant opportunities for the development of electronic commerce (e-commerce) systems. Home-based sellers use this technology to increase operational efficiency, reach a wider market, and make money from home through e-commerce platforms [1, 2]. However, home-based sellers face particular difficulties due to the wide range of products and many competitors on e-commerce platforms. Often, sellers new to online commerce do not adequately consider the selection of truly profitable products [3]. Therefore, if a company or seller achieves very high profits, it can be categorized as having good performance [4]. Consequently, a method is needed to assist home-based sellers in determining the most profitable products, both for the present and the future.

One commonly used approach is the Decision Support System, which helps decision makers choose the best alternative based on a number of criteria [5-7]. In this context, the concept of fuzzy logic becomes relevant because it can represent uncertainty and subjectivity in assessments. Introduced by Lotfi A. Zadeh in 1965, fuzzy logic is not strictly binary (true/false) but rather permits a value to exist within a range of 0 to 1 [8-10]. When making decisions involving information that is frequently unclear, ambiguous, or imprecise, this method is highly effective. The decision-making model's accuracy and realism can be enhanced by more flexibly describing the degree of importance or preference when criterion values are represented as fuzzy numbers, such as triangular fuzzy numbers (TFN) [9-12].

This study proposes two hybrid methods: FAHP-TOPSIS and FAHP-FTOPSIS. The FAHP-FTOPSIS method is an MCDM approach in which the Fuzzy Analytical Hierarchy Process (FAHP) is used to determine the relative weights of each criterion more flexibly by considering the uncertainty of the assessment [13-15]. Meanwhile, Fuzzy TOPSIS is used to rank alternatives based on their proximity to positive and negative ideal solutions in fuzzy space [16, 17]. As a comparison, the FAHP-TOPSIS method is used, which employs FAHP for weight determination and conventional TOPSIS for ranking, thereby enabling the evaluation of the feasibility and accuracy of the main method. Based on the research by Okfalisa et al. [18], five criteria are used in recommending online store products. The final results of this study will identify the method that produces the best ranking and determine the most suitable products to recommend to home-based sellers as a reference for selecting products to sell [19].

*Corresponding author.

II. Literature Review

A. Related Works

Previous studies have utilized fuzzy methods within the framework of Multi-Criteria Decision Making (MCDM).

A study by Ceren Erdin and Halil Emre Akba (2019) entitled "A Comparative Analysis of Fuzzy TOPSIS and Geographic Information Systems (GIS) for the Location Selection of Shopping Malls: A Case Study from Turkey", aimed to select strategic shopping mall locations in Turkey using the MCDM method with the Fuzzy TOPSIS and GIS approaches, and to compare the accuracy of each method [13]. In general, this study shows that combining fuzzy-based MCDM methods (e.g., Fuzzy-TOPSIS) with spatial analysis (e.g., GIS) makes location and alternative selection decisions more robust: ambiguous expert assessments can be handled, objective data is taken into account, and the results are more consistent and easier to visualize. As a result, the risk of wrong decisions is reduced, the planning process becomes more transparent and traceable, and this framework can be easily applied to other contexts (retail, logistics, energy, health).

In the study "Selection of Software Requirements using TOPSIS under Fuzzy Environment" by Mohd. Sadiq (2020), a method for selecting software requirements using Fuzzy TOPSIS was developed to address uncertainty and complexity in decision making [20]. In general, this research provides a structured and uncertainty-resistant method for prioritizing and selecting software requirements: linguistic expert assessments are converted into quantitative ratings using Fuzzy-TOPSIS (closeness coefficient), resulting in more objective and transparent decisions. This method is suitable for multi-stakeholders, addresses information ambiguity, and is relatively simple to calculate compared to several other techniques, as well as known to have minimal rank-reversal issues. The end result makes it easier for teams to select the most valuable "top-n" requirements and reduces the risk of misprioritization in development.

In a study conducted by Shamsuzzoha et al. [21] entitled "Application of the fuzzy TOPSIS framework for selecting complex projects", it offers a Fuzzy-TOPSIS-based selection framework to quickly identify the most complex projects, using eleven factors as a practical checklist, so that team allocation, budgeting, and support can be more targeted. The ranking results are proven to be stable through sensitivity testing, so decisions are not easily changed simply due to weight variations. As a result, the selection process becomes more objective, the risk of misplaced priorities is reduced, and project portfolio productivity increases while also filling a research gap due to its specific focus on selecting complex projects, rather than simply selecting general projects.

Previous studies have concluded that a research gap exists because studies on product selection with multiple criteria in e-commerce have not specifically addressed home-based sellers in Indonesia with large amounts of real data. The focus on indicators closely related to sellers (price, rating, sales, discounts, chat responses) is also still rare, as is a direct comparison between TOPSIS and Fuzzy TOPSIS with the same weighting basis and testing whether the results remain stable when there are small changes or when the criteria weights are shifted.

B. Product Selection

A crucial decision-making process is product selection: a process in which a variety of qualities and standards are evaluated to identify the best possible product. The application of cutting-edge techniques to deal with complexity and uncertainty in decision-making is emphasized in research on product selection. Fuzzy logic, hierarchical decision making, online reviews, multi-criteria analysis, and consumer expectations are some of the methods used [12, 22-25]. By taking consumer preferences, production efficiency, environmental impact, and technical viability into account, the objective is to maximize product selection [22]. The optimization method in the study by Tan et al. [23] focuses on determining the best product design that incorporates individual customer preferences and manufacturing constraints. Several studies on product selection show two main perspectives. First, the research by Gultom et al. [24] focuses on comprehending consumer preferences and using fuzzy sets to assist in decision-making, such as when prospective students are choosing a university. Second, the research by Aznag et al. [25] highlights how sellers should maximize revenue by optimizing product combinations while considering resource constraints.

C. MCDM

Multi-criteria decision-making (MCDM) is a method for evaluating and selecting the best choice from a number of options based on a number of factors. This approach is commonly used in areas such as selecting equipment, choosing technology, and evaluating service providers to improve performance, cost-efficiency, and other related aspects [26-28]. For these types of situations, multi-criteria decision methods (MCDMs) have been developed. These methods are also known for their formal structure and their ability to handle multiple conflicting goals and include different stakeholders in the decision-making process. Due to the wide range of problems they address, many multi-criteria decision methods have been created. One example is the Characteristic Objects METhod (COMET), which does not require the use of weights and is suitable for problems where rank reversal cannot happen [26]. Several MCDM methods can offer different solutions to the same problem. Variations in outcomes when using different computational approaches may result from factors such as the use of different weights for criteria, differences in the algorithms used, and the way targets are scaled [27, 29]. These observations highlight the importance of using a multi-method approach and techniques to fairly compare results in MCDM studies. Analyzing different MCDM methods shows that selecting the appropriate method and normalization technique is crucial. Almost any combination of a method and its parameters can lead to varied results [28]. Moreover, the increasing number of MCDM methods has created a need to compare the results they produce.

D. Home-Based Sellers

A home-based seller is typically a small-scale operation carried out from an individual's residence. These businesses contribute to household income and have a significant impact on the economy by using available resources, increasing

national output, generating employment opportunities, and supporting the livelihoods of individuals in urban areas. Despite these benefits, home-based sellers encounter challenges such as intense competition and limited sales [30-34]. Furthermore, various studies highlight that online home-based sellers offer a broad array of products and services, including food items, beauty products, clothing, artwork, handmade crafts, accounting services, consulting, freelance writing, virtual assistance, marketing, design, translation, web development, dropshipping, and caregiving [35-40].

There is a growing trend of home-based sellers participating in the digital marketplace or e-commerce, with a substantial portion of their revenue generated through online sales. The different categories of home-based sellers include: 1) home-based family businesses [35], 2) e-commerce home-based firms [37], 3) home-based sellers that focus on work-life balance [36], 4) diverse home-based sellers [37], and 5) women-owned home-based sellers [38].

In addition, several studies suggest that online home-based sellers sell a variety of products and services, such as food items, beauty products, clothing, artwork, homemade crafts, accountancy, consulting, freelance writing, virtual assistance, marketing, design, translation, web development, dropshipping, and caregiving [35-39].

E. FAHP

The fuzzy analytic hierarchy process (FAHP) is based on Zadeh's fuzzy set theory and Saaty's analytic hierarchy process (AHP) method for multi-criteria decision making (MCDM) [14, 41, 42]. The Hierarchy Analysis Process, or AHP, is the most commonly used. There are two main steps to using AHP in real life. The first step is to break a complicated problem down into a structured hierarchy. The second step is to determine the priority of each criterion to establish their importance at each level. The final stage involves evaluating the hierarchy based on the decision maker's preferences, where the weights from each matrix are calculated and normalized [28, 42]. One of the major advantages of this process is that it reduces a multi-dimensional problem to an equivalent unidimensional problem. This also facilitates producing a standard solution, which can be fused across several experts to come to a single agreed output. The method is flexible, and inconsistencies can be detected. It also provides tools for achieving consensus among experts. However, as noted in [43], the traditional AHP does not effectively address uncertainty in the evaluation of criterion importance. Because of this limitation, a fuzzy version of the method was adopted. This version enables the use of AHP in uncertain or ambiguous situations and has become increasingly accepted. It has been widely accepted and applied to address decision-making problems in various fields [44].

F. Fuzzy TOPSIS

The TOPSIS technique was proposed in 1981 by Hwang and Yoon and has been widely applied during the past several years. This strategy is beneficial for making decisions by comparison of alternative measures as a function of how close they are to an ideal point [28].

Fuzzy numbers are employed to handle uncertainty and subjective judgments in real-life situations. Intuitionistic Fuzzy Sets provide a more accurate representation of decision-makers' approvals, rejections, and uncertain or hesitant opinions [45].

G. Sesitivity Analysis

Sensitivity analysis aims to measure the accuracy of the validation values obtained in the case study and demonstrate the robustness of the hybrid model [46]. The determination of this sensitivity analysis is based on several smallest value ranges and a regression process. The following are various ways to determine sensitivity analysis [47]:

1) Mean Absolute Deviation (MAD): Mean Absolute Deviation (MAD) is a statistical measure that calculates the average absolute deviation of data from the central value, and is used to assess the extent of variation or instability in a system [48, 49].

2) Mean Squared Error (MSE): Mean Squared Error (MSE) is the bias (systematic deviation) of an estimate plus, statistically independent components that account for its variation [50], and thus can be used to indicate general accuracy levels of a predictive model.

3) Spearman's Correlation (SC): Spearman's Rank Correlation Coefficient is a non-parametric statistical measure used to assess the degree of monotonic relationship between two variables, especially when the data is expressed in the form of ranks or orders [51]. This coefficient measures the extent to which the relationship between two sets of ranks is consistent (rising or falling simultaneously) [52].

III. Methodology

A. Research Framework

A flowchart of this study is illustrated in Fig. 1.

The following is an explanation of the flowchart shown in Fig. 1:

1) Criteria and data setup

- Identify Criteria → Criteria Data: define the evaluation criteria (e.g., rating, price, sold, etc.) and prepare the criterion dataset.
- Alternative Data Identification / Collection / Processing → Alternative Data: identify the alternatives (products), collect raw data, clean/normalize as needed, and produce a ready-to-use alternatives dataset.

2) Weighting with FAHP (two parallel tracks)

- Establishing a Decision-Making Hierarchy → Compiling a Pair-Wise Matrix: build the decision hierarchy and create the pairwise comparison matrix of criteria.
- Determining the Level of Importance → CR $\leq$ 0.1: perform the consistency check (Consistency Ratio). If CR > 0.1, revise judgments; if CR $\leq$ 0.1, continue.
- Matrix Fuzzification → Weight Calculation → Weight Data: fuzzify judgments (e.g., triangular fuzzy numbers), compute criterion weights (defuzzify/normalize), and store them as Weight Data.

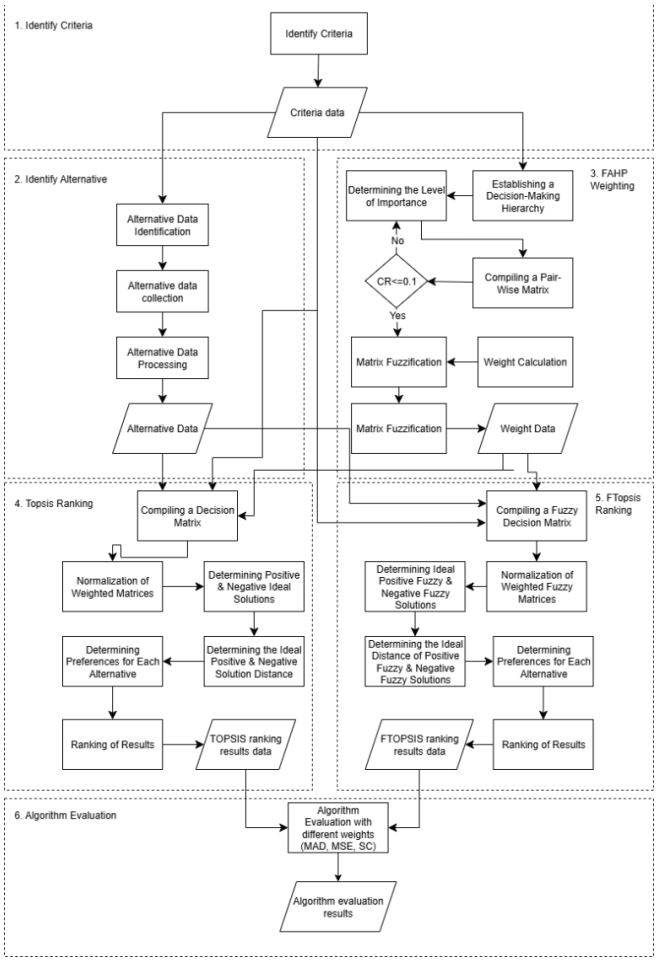


Fig. 1. Research method flowchart.

3) Decision matrix construction

- Compiling a Decision Matrix / Compiling a Fuzzy Decision Matrix: build the decision matrix (crisp for classical TOPSIS; fuzzy for FTOPSIS) that contains each alternative's performance on each criterion.

4) Ranking with TOPSIS (left branch)

- Normalizing the Weighted Matrices: normalize and apply weights to obtain the weighted normalized matrix.
- Determining the Positive and Negative Ideal Solutions: define the positive and negative ideal solutions.
- Determining the Ideal Positive and Negative Solution Distance: compute each alternative's distance to the positive and negative ideals.
- Determining the Preferences for Each Alternative → Ranking of Results → TOPSIS ranking results data: compute the preference score, rank the alternatives, and output TOPSIS ranking results.

5) Ranking with FTOPSIS (right branch)

- Normalizing the Weighted Fuzzy Matrices: perform fuzzy normalization and apply fuzzy weights.
- Determining the Ideal Positive and Negative Fuzzy Solutions: set the fuzzy positive/negative ideals.
- Determining the Ideal Distance of Positive Fuzzy and Negative Fuzzy Solutions: calculate distances in the fuzzy space (e.g., TFN distance).
- Determining the Preferences for Each Alternative → Ranking of Results → FTOPSIS ranking results data: (de)fuzzify preference scores, rank the alternatives, and output FTOPSIS ranking results.

6) Sensitivity evaluation across weight sets

- Algorithm evaluation with different weights (MAD, MSE, SC): test the stability of both algorithms under weight changes using three metrics:
  - MAD (Mean Absolute Deviation): average absolute rank difference vs. the baseline (smaller = more stable).
  - MSE (Mean Square Error): average squared rank difference (more sensitive to large shifts).
  - SC (Spearman Correlation): rank-order consistency (closer to 1 = highly consistent).
- Algorithm evaluation results: summarize which algorithm is more stable/robust under the tested weight profiles.

B. Data Collection and Criteria

1) Dataset source: This study uses data on food and beverage products from various stores on the Shopee Indonesia e-marketplace, collected using web scraping methods, amounting to 10,000 data points.

The criteria used in the study were obtained from the questionnaire results to determine the importance of each criterion. Referring to previous studies and adding two new criteria, which are subscribers and followers. The criteria used can be seen in Table I.

TABLE I. CRITERIA FOR SELECTING THE BEST PRODUCTS

Criteria	Description	Type
Rating	Product evaluation by consumers	Benefit
Price	Selling price of products	Cost
Sold	Number of units sold of a product	Benefit
Discount	Discounts on products	Benefit
Response Chat	Percentage of conversations responded to by sellers	Benefit

C. Fuzzy Analytic Hierarchy Process (FAHP

Here are the steps for weighting FAHP:

1) Conversion of multi-criteria decision-making problems into hierarchical models.

2) Identification of the importance of the pair-wise comparison matrix in determining the importance of criteria on the Saaty comparison scale is given in Table II.

TABLE II. INTENSITY OF INTEREST

Verbal scale	Numerical scale
1	Both elements are equal
3	A little more important
5	Most important
7	Much more important
9	Absolutely more important
2,4,6, and 8	Intermediate value
invers	$aij = 1/aij$

3) Normalization of the weight of each criterion importance value and find its eigenvector value.

4) Calculation of the consistency ratio using the index ratio to check the consistency of the weight values is given in Eq. (1):

$$CI = \frac{\lambda max - n}{RI*(n-1)} \quad (1)$$

If the CR value is ≤ (less than or equal to) 0.1, then the criteria values are consistent. However, if not, a re-evaluation is necessary to ensure the consistency of the criteria values.

5) Fuzzification of Pair-Wise Comparison Matrix into TFN Fuzzy Scale is given in Table III.

TABLE III. PAIR-WISE FUZZY SCALE

Intensity of importance	Fuzzy Scale (l,m,u)
1	(1,1,1)
2	(1,2,3)
3	(2,3,4)
4	(3,4,5)
5	(4,5,6)
6	(5,6,7)
7	(6,7,8)
8	(7,8,9)
9	(8,9,10)
invers	$(l,m,u)_{ij} = (1/u, 1/m, 1/l)_{ij}$

6) Fuzzy weight calculation using Fuzzy Geometrix Mean [Eq. (2)]:

$$\widetilde{w}_i = \tilde{r}_i \otimes (\tilde{r}_1 \oplus \tilde{r}_2 \oplus \dots \oplus \tilde{r}_n)^{\frac{1}{n}} \quad (2)$$

7) Defuzzification of fuzzy weights using the Center Of Area method [Eq. (3)]:

$$w_i = (lw_i + mw_i + uw_i)/3 \quad (3)$$

8) Weight normalization [Eq. (4)]:

$$w_r = \frac{w_i}{\sum_{i=1}^{n} w_i} \quad (4)$$

D. Technique for Order Preference by Similarity to Ideal Solution (TOPSIS)

The steps of TOPSIS are as follows:

1) Determination of the normalization of the decision matrix.

The normalized weight value rij is calculated using the following Eq. (5):

$$r_{ij} = \frac{Xij}{\sqrt{\sum_{i=1}^{m} X_{ij}^2}} \quad (5)$$

2) Determination of the normalized weights of the decision matrix.

The normalized weight value yij is calculated using the following Eq. (6):

$$yij = Wij\, x\, rij \quad (6)$$

provided that:

$$A+ = (y_1^+, y_2^+, \dots, Y_N^+)$$
$$A- = (y_1^-, y_2^-, \dots, Y_N^-)$$

Description:

$y+ = (max\ _i y_{ij}$;if j is an alternative category of benefits)

$y- = (min\ _i y_{ij}$; if j is an alternative category of Costs)

With a value of j = 1, 2, …n

3) Calculation of the distance between the Ai alternative and the positive ideal solution using the following Eq. (7):

$$D_i^+ = \sqrt{\sum_{j=1}^{n} (y_j^+ - y_{ij})^2} \quad (7)$$

4) Calculation of the distance between the Ai alternative and the ideal solution using Eq. (8):

$$D_i^- = \sqrt{\sum_{j=1}^{n} (y_{ij} + y_j^-)^2} \quad (8)$$

5) Calculation of the preference value for each alternative (Vi) using Eq. (9):

$$V_i = \frac{D_i^-}{D_i^- + D_I^+} \quad (9)$$

E. Fuzzy Technique for Order Preference by Similarity to Ideal Solution (FTOPSIS)

The steps of Fuzzy TOPSIS are as follows:

1) Fuzzyfication of the decision matrix into TFN fuzzy values becomes three values: low, medium, and upper.

2) Determination of the normalization of the decision matrix. The normalized weight values rij are calculated using the following equation:

Benefit Criteria [Eq. (10)]:

$$\check{r}_{ij} = \left(\frac{a_{ij}}{c_i^{MAX}} + \frac{b_{ij}}{b_i^{MAX}} + \frac{c_{ij}}{a_i^{MAX}}\right) \quad (10)$$

Cost Criteria [Eq. (11)]:

$$\check{r}_{ij} = \left(\frac{c_i^{MAX}}{a_{ij}} + \frac{b_i^{MAX}}{b_{ij}} + \frac{a_i^{MAX}}{c_{ij}}\right) \quad (11)$$

provided that:

$$A+ = \left(\max a_{ij} \max b_{ij}, \max c_{ij}\right)$$

$$A- = \left(\min a_{ij} \min b_{ij}, \min c_{ij}\right)$$

3) Calculation of the distance between the Ai alternative and the positive ideal solution using Eq. (12):

$$D_i^+ = \sum_{i=j}^{n} \sqrt{\frac{1}{3}\{(a_{ij} - A_i^+) + \left(b_{ij} - A_i^+\right) + (c_{ij} - A_i^+)\}} \quad (12)$$

4) Calculation of the distance between the Ai alternative and the negative ideal solution using Eq. (13):

$$D_i^- = \sum_{i=j}^{n} \sqrt{\frac{1}{3}\{(a_{ij} - A_i^-) + \left(b_{ij} - A_i^-\right) + (c_{ij} - A_i^-)\}} \quad (13)$$

5) Calculation of the preference value for each alternative (Vi) using Eq. (14):

$$V_i = \frac{D_1^-}{D_1^- + D_I^+} \quad (14)$$

F. Sensitivity Analysis

The following are the equations for sensitivity analysis.

1) Mean Absolute Deviation (MAD)

The following is the equation for MAD [Eq. (15)]:

$$MAD = \frac{1}{n}\sum_{i=1}^{n} |xij - \text{YJ}| \quad (15)$$

2) Mean Squared Error (MSE)

The following is the equation for MSE [Eq. (16)]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (xij - yj)^2 \quad (16)$$

3) Spearman's Correlation (SC)

The following is the equation for SC [Eq. (17)]:

$$SC = r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \quad (17)$$

IV. Results and Discussion

A. Criteria Weight Calculation Using FAHP

The first step is to convert the multi-criteria decision-making problem into a hierarchical model consisting of objectives, criteria, sub-criteria (if any), and alternatives. The following is a decision hierarchy illustrated in Fig. 2.

The hierarchical model in this study consists of three levels: objectives, criteria, and alternatives. The first level is the objective, which is to determine the most potential product to sell for home-based sellers. The second level consists of five criteria: rating, price, sold, discount, and chat response. The third level contains product alternatives to be evaluated: Product 1 to Product N.

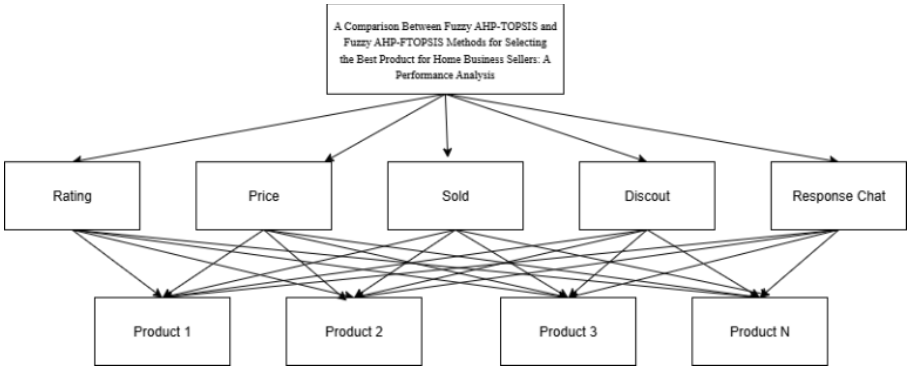


Fig. 2. Decision hierarchy.

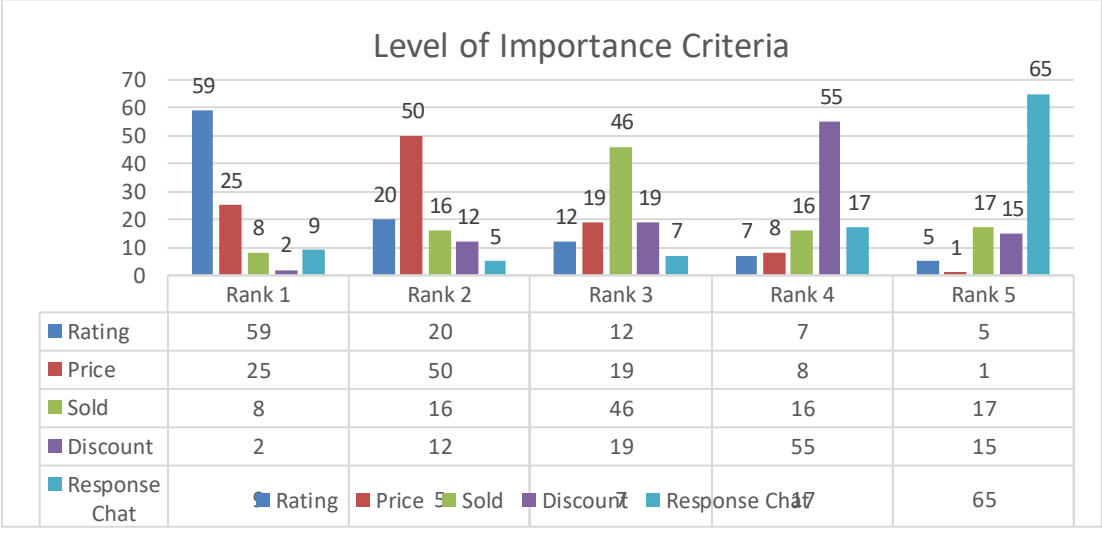


Fig. 3. Level of importance criteria.

After establishing the hierarchy, the next step was to determine the level of importance of each criterion. This was done by distributing questionnaires to sellers and successfully obtaining 103 respondents. The results of determining the level of importance are shown in Fig. 3.

Once the criteria were known, pair-wise comparisons were then conducted for each criterion, as in Table IV.

TABLE IV. MATRIX PAIR-WISE COMPARISON

	Rating	Price	Sold	Discount	Response Chat
Rating	1	2	3	4	5
Price	$\frac{1}{2}$	1	2	3	4
Sold	$\frac{1}{3}$	$\frac{1}{2}$	1	2	3
Discount	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	1	2
Response Chat	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{1}{2}$	1

The results of the weights (eigenvectors) of each criterion from the pair-wise comparison results can be used to calculate the best product selection, as shown in Table V.

TABLE V. EIGENVECTOR

Criteria	Eigenvector
Rating	0.416
Price	0.262
Sold	0.161
Discount	0.099
Response Chat	0.062

Then, the eigenvector is evaluated by determining the Consistency Ratio (CR) using Eq. (1). The Consistency Ratio obtained is 0.022, which is less than 0.1, so the matrix obtained previously is consistent and does not need to be corrected.

After the matrix is consistent, fuzzification is performed on the pair-wise matrix according to the FAHP matrix scale, as in Table III. The following are the fuzzification results as shown in Table VI.

TABLE VI. FUZZY MATRIX PAIR-WISE

	Rating	Price	Sold	Discount	RC
Rating	(1;1;1)	(1;2;3)	(2;3;4)	(3;4;5)	(4;5;6)
Price	$\left(\frac{1}{3};\frac{1}{2};1\right)$	(1;1;1)	(1;2;3)	(2;3;4)	(3;4;5)
Sold	$\left(\frac{1}{4};\frac{1}{3};\frac{1}{2}\right)$	$\left(\frac{1}{3};\frac{1}{2};1\right)$	(1;1;1)	(1;2;3)	(2;3;4)
Discount	$\left(\frac{1}{5};\frac{1}{4};\frac{1}{3}\right)$	$\left(\frac{1}{4};\frac{1}{3};\frac{1}{2}\right)$	$\left(\frac{1}{3};\frac{1}{2};1\right)$	(1;1;1)	(1;2;3)
RC	$\left(\frac{1}{6};\frac{1}{5};\frac{1}{4}\right)$	$\left(\frac{1}{5};\frac{1}{4};\frac{1}{3}\right)$	$\left(\frac{1}{4};\frac{1}{3};\frac{1}{2}\right)$	$\left(\frac{1}{3};\frac{1}{2};1\right)$	(1;1;1)

After the matrix has been fuzzified, the fuzzy weight will be calculated using Eq. (2), and the results of the calculation are shown in Table VII.

TABLE VII. FUZZY WEIGHT

Criteria	L	M	U
Rating	0,421	0,417	0,389
Price	0,256	0,263	0,272
Sold	0,156	0,160	0,171
Discount	0,098	0,097	0,104
Response Chat	0,069	0,062	0,063

Next, defuzzification of the fuzzy weights from the fuzzy weight calculations performed using Eq. (3), and the weights were normalized using Eq. (4). The results can be seen in Table VIII.

TABLE VIII. CRIPS WEIGHT

Criteria	Weight
Rating	0.409
Price	0.264
Sold	0.163
Discount	0.100
Response Chat	0.065
CR	0.015

B. Product Ranking using TOPSIS

After obtaining the weight of each criterion, the calculation of the best product ranking was carried out using 10,000 (ten thousand) alternatives obtained from the Shopee Indonesia website, as shown in Table IX.

TABLE IX. DATASET TOPSIS

A/C	C1	C2	C3	C4	C5
A1	2	5000	10	0	14
A2	4	55000	5	0	82
A3	3	88000	2	5	50
A4	5	11000	9	0	79
A5	5	17999	10	0	41
A6	5	18900	31	0	38
A7	5	9500	2	0	96
A8	5	75900	64	0	91
A9	5	39999	125	0	66
A10	4	13000	83	30	94
…….					
A10000	5	4750	8	0	89

Next, calculations were performed using the TOPSIS method with Eq. (6) for weight normalization, Eq. (7) and Eq. (8) for negative and positive ideal values, and Eq. (9) to find the utility value of each alternative, as in Table X.

After that, the scores were calculated and ranked from the highest to the lowest. The following are the top 10 rankings as shown in Table XI.

TABLE X. UTILITY VALUE

A/C	C1	C2	C3	C4	C5
A1	0,00029	0,00128	0,00007	0,00000	0,00013
A2	0,00322	0,00256	0,00003	0,00000	0,00074
A3	0,00515	0,00192	0,00001	0,00043	0,00045
A4	0,00064	0,00320	0,00006	0,00000	0,00071
A5	0,00105	0,00301	0,00007	0,00000	0,00037
A6	0,00111	0,00308	0,00021	0,00000	0,00034
A7	0,00056	0,00320	0,00001	0,00000	0,00087
A8	0,00444	0,00308	0,00044	0,00000	0,00082
A9	0,00234	0,00308	0,00085	0,00000	0,00060
A10	0,00076	0,00282	0,00057	0,00258	0,00085
.....					
A10000	0,00028	0,00320	0,00005	0,00000	0,00080

TABLE XI. TOPSIS RANKING RESULTS

Rank	Alternative	Score
1	A5948	0,945977
2	A5950	0,897357
3	A6375	0,775835
4	A6625	0,705012
5	A3358	0,668346
6	A2667	0,649551
7	A3930	0,563599
8	A5312	0,498368
9	A776	0,423548
10	A6521	0,415185

C. Product Ranking Using FTOPSIS

The calculation using the FTOPSIS method was performed using fuzzy weights with three values from FAHP. The dataset used the data in Table IX, which was converted into Fuzzy TFN form in Table XII.

Next, calculations were performed using the FTOPSIS method with Eq. (10) for normalization if the criterion was Benefit and Eq. (11) if the criterion was Cost, as well as Eq. (12) and Eq. (13) for negative Fuzzy and positive Fuzzy ideal values. Then, the score was calculated using Eq. (14). The final results can be seen in Table XIII.

D. Graph of Ranking Changes

To evaluate the impact of weight variations on the ranking results, a comparative analysis was conducted using both Fuzzy TOPSIS and classical TOPSIS methods. The results are presented in the following figures, which illustrate how the rankings of alternatives change under different weight scenarios.

The first graph in Fig. 4 shows the ranking changes of alternatives under different weight scenarios using the Fuzzy TOPSIS method. Most alternatives show relatively stable rankings with minor shifts, but A919 and A5774 demonstrate sharp fluctuations, especially under the Second Weight scenario, where their ranks rise drastically. This indicates that Fuzzy TOPSIS is more sensitive to weight variations, particularly for certain alternatives, and may produce significant rank changes when the weights are adjusted.

TABLE XII. DATA FTOPSIS

A/C	C1	C2	C3	C4	C5
A1	(1.8;2.0; 2.2)	(4500;5000;5500)	(9.0;10;11)	(0;0;0)	(12.6;14; 15.4)
A2	(3.6;4.0; 4.4)	(49500;55000; 60500)	(4.5;5;5.5)	(0;0;0)	(73.8;82; 90.2)
A3	(2.7;3.0; 3.3)	(79200;88000; 96800)	(1.8;2;2.2)	(4.5;5; 5.5)	(45;50;55)
A4	(4.5;5.0; 5.5)	(9900;11000;12100)	(8.1;9;9.9)	(0;0;0)	(71.1;79; 86.9)
A5	(4.2;4.7; 5.2)	(16199;17999; 19799)	(9.0;10;11)	(0;0;0)	(36.9;41; 45.1)
A6	(4.3;4.8; 5.3)	(17010;18900; 20790)	(27.9;31;34.1)	(0;0;0)	(34.2;38; 41.8)
A7	(4.5;5.0; 5.5)	(8550;9500;10450)	(1.8;2;2.2)	(0;0;0)	(86.4;96; 106)
A8	(4.3;4.8; 5.3)	(68310;75900; 83490)	(57.6;64;70.4)	(0;0;0)	(81.9;91; 100)
A9	(4.3;4.8; 5.3)	(35999;39999; 43999)	(112.5;125 ;138)	(0;0;0)	(59.4;66; 72.6)
A10	(4.0;4.4; 4.8)	(11700;13000; 14300)	(74.7;83;91.3)	(27;30; 33)	(84.6;94; 103)
.....					
A10000	(4.5;5.0; 5.5)	(4275;4750;5225)	(7.2;8;8.8)	(0;0;0)	(80.1;89; 97.9)

TABLE XIII. FTOPSIS RANKING RESULTS

Rank	Alternative	Score
1	A5948	0,67910
2	A5950	0,63945
3	A6375	0,60883
4	A771	0,56562
5	A2667	0,56104
6	A2025	0,52758
7	A2384	0,52582
8	A5774	0,52290
9	A919	0,52283
10	A1623	0,52274

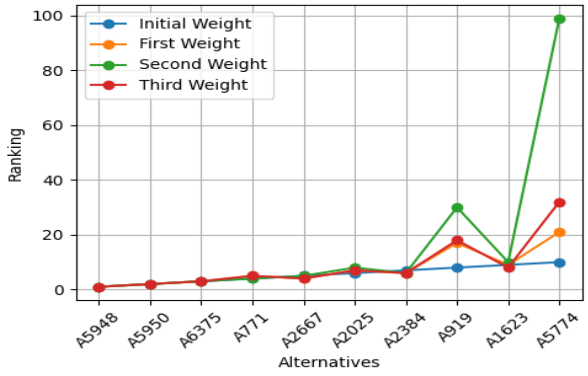


Fig. 4. FTOPSIS ranking change chart.

Fig. 5. TOPSIS ranking change chart.

The second graph in Fig. 5 displays the ranking results under different weight scenarios using the classical TOPSIS method. In this case, the rankings remain perfectly consistent across all scenarios, with all lines overlapping into a straight diagonal pattern. This demonstrates that TOPSIS yields a highly stable ranking outcome that is unaffected by weight modifications, showing robustness in preserving rank order despite changes in weights.

Together, the two graphs highlight a contrast: Fuzzy TOPSIS provides more flexibility but introduces higher sensitivity to weight changes, while TOPSIS delivers more stable and robust results, maintaining the same ranking across different scenarios.

E. Sensitivity Analysis

Sensitivity analysis was conducted by re-ranking alternatives using three different FAHP weight datasets and comparing the rankings from the main weights with those from the experimental weights to assess the impact of weight changes on alternative rankings, using Mean Absolute Deviation (MAD), Mean Square Error (MSE), and Spearman Correlation (SC).

1) MAD: The following are the MAD results of the two methods used, as shown in Fig. 6.

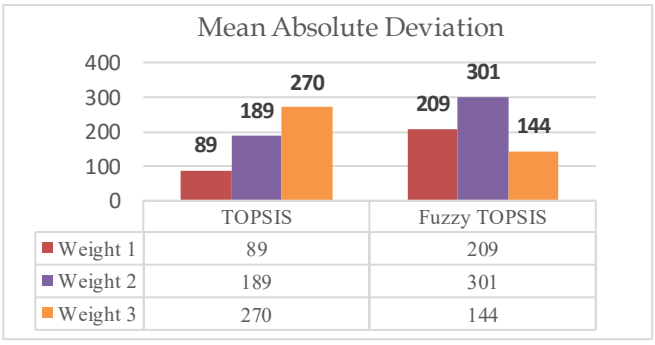


Fig. 6. Mean absolute deviation graph.

The results of the Mean Absolute Deviation (MAD) calculation in Fig. 6 show that TOPSIS has the highest ranking stability in Weight 1 (89) but is sensitive in Weight 3 (270). Conversely, Fuzzy TOPSIS is relatively unstable in Weight 1 (209) and most sensitive in Weight 2 (301), but shows the best stability in Weight 3 (144). In general, TOPSIS is more stable in the initial weights, while Fuzzy TOPSIS performs better in certain weights, but with greater fluctuations.

2) MSE: The following are the MSE results of the two methods used, as shown in Fig. 7.

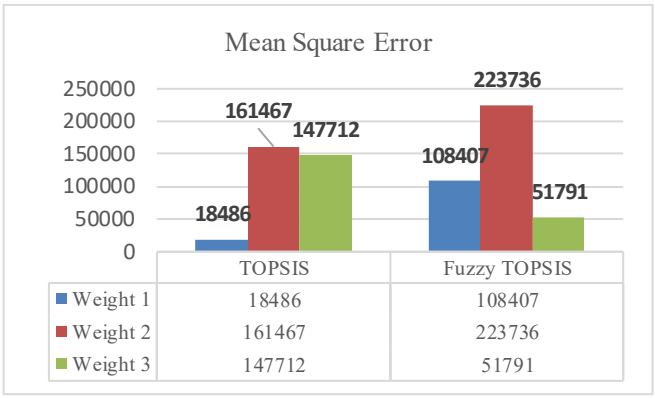


Fig. 7. Mean square error graph.

The Mean Square Error (MSE) results in Fig. 5 show that TOPSIS has the smallest error in Weight 1 (18.486) and increases significantly in Weight 2 (161.467) and Weight 3 (147.712). Conversely, Fuzzy TOPSIS has the lowest MSE for Weight 3 (51.791) but the highest for Weight 2 (223.736). In general, TOPSIS is more stable at the initial weights, while Fuzzy TOPSIS performs better for Weight 3, despite having larger error fluctuations.

3) SC: The following are the SC results of the two methods used, as shown in Fig. 8.

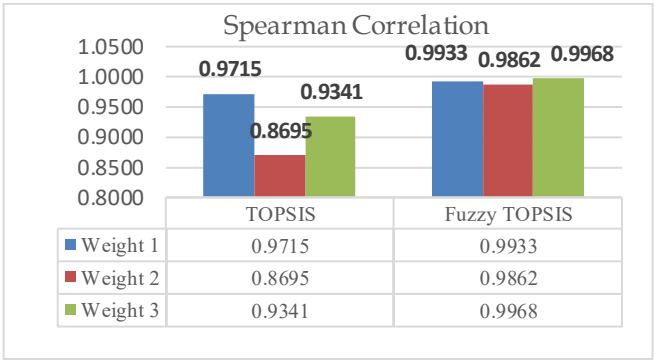


Fig. 8. Spearman correlation graph.

The Spearman Correlation results in Fig. 8 show that Fuzzy TOPSIS has a very high and consistent correlation in all weight scenarios: 0.9933 (Weight 1), 0.9862 (Weight 2), and 0.9968 (Weight 3). Meanwhile, TOPSIS recorded the highest correlation in Weight 1 (0.9715) and the lowest in Weight 2 (0.8695), with a moderate value in Weight 3 (0.9341). Overall, Fuzzy TOPSIS demonstrated better ranking stability compared to TOPSIS across all weight scenarios. Overall, Fuzzy TOPSIS demonstrated better ranking stability compared to TOPSIS across all weight scenarios.

Based on the results of testing using three evaluation metrics, namely Mean Absolute Deviation (MAD), Mean Square Error (MSE), and Spearman Correlation (SC), it was

found that TOPSIS showed the best stability at the initial weight (MAD = 89; MSE = 18.486; SC = 0.9715), while Fuzzy TOPSIS outperformed it at Weight 3 (MAD = 144; MSE = 51.791; SC = 0.9968) and maintained consistently high correlation values across all weight scenarios. However, Fuzzy TOPSIS tends to experience greater error fluctuations in Weight 2. These findings indicate that the selection of the appropriate method and weight settings significantly influences the stability and accuracy of the ranking results.

Based on sensitivity analysis using MAD, MSE, and Spearman's Correlation (SC), it can be concluded that at the base weight (Weight 1), TOPSIS shows the best ranking stability—characterized by relatively low mean absolute deviation and mean square error as well as high ranking correlation with the reference. However, when the weights are shifted (particularly in the Weight 3 pattern), Fuzzy TOPSIS (FTOPSIS) is superior in maintaining consistency in the order, as reflected in the SC value that is very close to 1, even though the error profile increased in the Weight 2 configuration.

V. Conclusion

This study fills the gap by constructing a simple, field data-based model, comparing two approaches (FAHP–TOPSIS and FAHP–FTOPSIS), and then testing the stability of the results in scenarios of minor changes and significant weight changes. From this, practical guidelines that are easy to follow are compiled: Fuzzy TOPSIS is more appropriate when the data tends to be vague but the changes are small, while TOPSIS is more suitable when the focus of the assessment shifts significantly. The process is easy to repeat, and the consistent peak results in both approaches show that the recommendations produced are reliable.

The results of the comparative ranking analysis using the FAHP-TOPSIS and FAHP-FTOPSIS hybrid methods to support decision making in selecting potential products for home-based sellers. The alternative ranking results show that the first-ranked alternative is the same in both the FAHP-TOPSIS and FAHP-FTOPSIS methods: alternative A5948; however, the scores are different. In the FAHP-TOPSIS results, the score for this alternative is 0.945977, while in the FAHP-FTOPSIS results, this alternative receives a score of 0.67910. This proves that the addition of the Fuzzy approach to TOPSIS does not affect the product ranking results but does affect the score values of the product alternatives.

The sensitivity test results show that, FAHP-TOPSIS is stronger in dealing with changes in criterion weights, with the best results for weight one. Mean Absolute Deviation 89, Mean Square Error 18.486, and Spearman Correlation 0.972. Conversely, FAHP-FTOPSIS shows superiority at weight three—Mean Absolute Deviation 144, Mean Square Error 51.791, and Spearman Correlation 0.997, though its performance is more fluctuating at other weights. Practically, for baseline weighting scenarios in product selection, TOPSIS is the most stable choice; while for scenarios with weight variations (e.g., shifting seller preferences), FTOPSIS offers better ranking robustness, with the caveat that controlling error variance at certain weights remains necessary.

In short, FAHP-FTOPSIS is the safer choice for messy, real-world settings with uncertain or shifting preferences because it keeps rankings more stable. FAHP-TOPSIS can be attractive when decision makers have crisp, well-defined weights and you mainly care about minimizing numerical error. So, use FAHP-FTOPSIS as the default under ambiguity; switch to FAHP-TOPSIS when preferences are clear and consistency is less of a concern.

These findings can serve as a practical guideline for sellers. When preferences are uncertain or may shift over time as is common in dynamic marketplaces, FAHP-FTOPSIS should be the default because it keeps rankings stable and reliable under ambiguity. When priorities are crisp and well agreed upon, FAHP-TOPSIS can be used to emphasize minimal numerical error. In both cases, the resulting rankings are actionable: sellers can use them to shortlist products, allocate budget, and schedule promotions with greater confidence. We also recommend reviewing weights periodically (e.g., when market conditions or campaign goals change) and re-running the analysis so that the ranking continues to reflect current strategy and risk tolerance.

References

[1] D. Reuschke and C. Mason, "The engagement of home-based businesses in the digital economy," Futures, vol. 135, p. 102542, 2022.

[2] R. J. Nathan, V. Victor, C. L. Gan, and S. Kot, "Electronic commerce for home-based businesses in emerging and developed economy," Eurasian Business Review, vol. 9, no. 4, pp. 463-483, 2019.

[3] X. Pu, S. Sun, and J. Shao, "Direct selling, reselling, or agency selling? Manufacturer's online distribution strategies and their impact," International Journal of Electronic Commerce, vol. 24, no. 2, pp. 232-254, 2020.

[4] B. Cuellar-Fernández, Y. Fuertes-Callén, and C. Serrano-Cinca, "Survival of e-commerce entrepreneurs: The importance of brick-and-click and internationalization strategies," Electronic Commerce Research and Applications, vol. 46, p. 101035, 2021/03/01/ 2021, doi: https://doi.org/10.1016/j.elerap.2021.101035.

[5] M. Hosoya, H. Yamaura, S. Nakamura, M. Nakamura, E. Takamatsu, and Y. Kitaide, "Does the pop-out make an effect in the product selection of signage vending machine?," in Human-Computer Interaction–INTERACT 2019: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part II 17, 2019: Springer, pp. 24-32.

[6] S. Ma and Y. Liang, "Human Resource Management Decision System University Based on Decision Model," Mathematical Problems in Engineering, vol. 2023, no. 1, p. 6212569, 2023.

[7] A. Bączkiewicz, B. Kizielewicz, A. Shekhovtsov, J. Wątróbski, and W. Sałabun, "Methodical aspects of MCDM based E-commerce recommender system," Journal of Theoretical and Applied Electronic Commerce Research, vol. 16, no. 6, pp. 2192-2229, 2021.

[8] D. Jhodkar, A. Khan, and K. Gupta, "Fuzzy-MOORA based optimization of machining parameters for machinability enhancement of titanium," 2021.

[9] G. Chen, T. Pham, and N. Boustany, "Introduction to Fuzzy Sets, Fuzzy Logic, and Fuzzy Control Systems," Applied Mechanics Reviews - APPL MECH REV, vol. 54, 11/01 2001, doi: 10.1115/1.1421114.

[10] A.-T. Nguyen, T. Taniguchi, L. Eciolaza, V. Campos, R. Palhares, and M. Sugeno, "Fuzzy control systems: Past, present and future," IEEE Computational Intelligence Magazine, vol. 14, no. 1, pp. 56-68, 2019.

[11] F. Wang, "Preference degree of triangular fuzzy numbers and its application to multi-attribute group decision making," Expert Systems with Applications, vol. 178, p. 114982, 2021/09/15/ 2021, doi: https://doi.org/10.1016/j.eswa.2021.114982.

[12] M. Borza and A. S. Rambely, "A new method to solve multi-objective linear fractional problems," Fuzzy information and engineering, vol. 13, no. 3, pp. 323-334, 2021.

[13] C. Erdin and H. E. Akbaş, "A comparative analysis of fuzzy TOPSIS and geographic information systems (GIS) for the location selection of shopping malls: a case study from Turkey," Sustainability, vol. 11, no. 14, p. 3837, 2019.

[14] P. Data, "Fuzzy analytical hierarchy process (FAHP) using geometric mean method to select best processing framework adequate to big data," Journal of Theoretical and Applied Information Technology, vol. 99, no. 1, pp. 207-226, 2021.

[15] Y. Liu and X. Zhang, "Evaluating the Undergraduate Course based on a Fuzzy AHP-FIS Model," International Journal of Modern Education & Computer Science, vol. 12, no. 6, 2020.

[16] F. Kutlu Gündoğdu and C. Kahraman, "A novel fuzzy TOPSIS method using emerging interval-valued spherical fuzzy sets," Engineering Applications of Artificial Intelligence, vol. 85, pp. 307-323, 2019/10/01/ 2019, doi: https://doi.org/10.1016/j.engappai.2019.06.003.

[17] A. Memari, A. Dargi, M. R. A. Jokar, R. Ahmad, and A. R. A. Rahim, "Sustainable supplier selection: A multi-criteria intuitionistic fuzzy TOPSIS method," Journal of manufacturing systems, vol. 50, pp. 9-24, 2019.

[18] D. U. Iswavigra, "Online Shop Recommendations: Decision Support System Based on Multi-Objective Optimization on the Basis of Ratio Analysis," in 2020 8th International Conference on Cyber and IT Service Management (CITSM), 2020: IEEE, pp. 1-6.

[19] S. M. AL-Ghuribi, S. A. Mohd Noah, S. Tiun, M. A. Mohammed, and N. I. Y. Saat, "Combining review elements for modelling various multi-criteria collaborative recommendation models," Journal of Big Data, vol. 12, no. 1, p. 160, 2025.

[20] M. Sadiq, K. Shadab, and C. W. and Mohammad, "Selection of software requirements using TOPSIS under fuzzy environment," International Journal of Computers and Applications, vol. 44, no. 6, pp. 503-512, 2022/06/03 2022, doi: 10.1080/1206212X.2020.1820689.

[21] A. Shamsuzzoha, S. Piya, and M. Shamsuzzaman, "Application of fuzzy TOPSIS framework for selecting complex project in a case company. J Glob Oper Strateg 14 (3): 528–566," ed, 2021.

[22] G. Zhang and H. Qiu, "Competitive product identification and sales forecast based on consumer reviews," Mathematical Problems in Engineering, vol. 2021, no. 1, p. 2370692, 2021.

[23] C. Tan, H. Chung, K. Barton, S. J. Hu, and T. Freiheit, "Incorporating customer personalization preferences in open product architecture design," Journal of Manufacturing Systems, vol. 56, pp. 72-83, 2020.

[24] P. Gultom and H. Mulyono, "Fuzzy Sets Applied for Product Selection based on Customer Preferences," 2022.

[25] A. Aznag, V. Goyal, and N. Perivier, "MNL-Bandit with Knapsacks: a near-optimal algorithm," arXiv preprint arXiv:2106.01135, 2021.

[26] R. Robina-Ramírez, A. Chamorro-Mera, and L. Moreno-Luna, "Organic and online attributes for buying and selling agricultural products in the e-marketplace in Spain," Electronic Commerce Research and Applications, vol. 42, p. 100992, 2020.

[27] B. Agarski, M. Hadzistevic, I. Budak, S. Moraca, and D. Vukelic, "Comparison of approaches to weighting of multiple criteria for selecting equipment to optimize performance and safety," International Journal of Occupational Safety and Ergonomics, vol. 25, no. 2, pp. 228-240, 2019.

[28] O. Boutkhoum, M. Hanine, T. Agouti, and A. Tikniouine, "A decision-making approach based on fuzzy AHP-TOPSIS methodology for selecting the appropriate cloud solution to manage big data projects," International Journal of System Assurance Engineering and Management, vol. 8, no. Suppl 2, pp. 1237-1253, 2017.

[29] M. Shantal, Z. Othman, and A. A. Bakar, "A novel approach for data feature weighting using correlation coefficients and min–max normalization," Symmetry, vol. 15, no. 12, p. 2185, 2023.

[30] G. Haynes et al., "Family business research: Reviewing the past, contemplating the future," Journal of Family and Economic Issues, vol. 42, no. Suppl 1, pp. 70-83, 2021.

[31] M. N. Anwar and E. Daniel, "The role of entrepreneur-venture fit in online home-based entrepreneurship: A systematic literature review," Journal of Enterprising Culture, vol. 24, no. 04, pp. 419-451, 2016.

[32] K. Kane and W. A. Clark, "Mapping the landscape of urban work: Home-based businesses and the built environment," Environment and Planning A: Economy and Space, vol. 51, no. 2, pp. 323-350, 2019.

[33] A. Ekhlassi, F. Fazli, and E. Alinaghian, "Marketing and Branding for Iranian Home-Based Businesses," in Iranian Entrepreneurship: Deciphering the Entrepreneurial Ecosystem in Iran and in the Iranian Diaspora: Springer, 2017, pp. 343-359.

[34] W. Tyas and J. Hutama, "Strategy and innovation of home based enterprises for local development in the 4.0 era: a bibliographic study," in IOP Conference Series: Earth and Environmental Science, 2021, vol. 673, no. 1: IOP Publishing, p. 012054.

[35] M. Madanoglu, E. Memili, and A. De Massis, "Home-based family firms, spousal ownership and business exit: a transaction cost perspective," Small business economics, vol. 54, no. 4, pp. 991-1006, 2020.

[36] R. Gorjifard and J. Crawford, "Working from home: Impact on wellbeing and work-life balance," New Zealand Journal of Employment Relations, vol. 46, no. 2, pp. 64-78, 2021.

[37] L. Philip and F. Williams, "Remote rural home based businesses and digital inequalities: Understanding needs and expectations in a digitally underserved community," Journal of Rural Studies, vol. 68, pp. 306-318, 2019.

[38] H. A. Seshie-Nasser and A. D. Oduro, "Women-owned businesses and household welfare," International Journal of Gender and Entrepreneurship, vol. 10, no. 4, pp. 310-331, 2018.

[39] T. M. Osaili, A. A. Al-Nabulsi, and T. M. Al-Jaberi, "Food safety knowledge, attitudes, and practices among female food handlers engaged in home-based online food businesses in Jordan during the COVID-19 pandemic," Heliyon, vol. 8, no. 8, 2022.

[40] A. N. Azmi, A. A. Bakar, Z. A. Othman, and N. S. Sani, "Mining Association Rules to Determine the Overspending Behavior Among Low Income Households in Malaysia," Asia-Pacific Journal of Information Technology & Multimedia, vol. 14, no. 1, 2025.

[41] T. L. Saaty, "Optimization by the analytic hierarchy process," 1979.

[42] M. Bakır and Ö. Atalık, "Application of Fuzzy AHP and Fuzzy MARCOS Approach for the Evaluation of E-Service Quality in the Airline Industry," Decision Making: Applications in Management and Engineering, vol. 4, no. 1, pp. 127-152, 03/13 2021, doi: 10.31181/dmame2104127b.

[43] M. T. Al Fozaie and H. Wahid, "A guide to integrating expert opinion and fuzzy AHP when generating weights for composite indices," Advances in Fuzzy Systems, vol. 2022, no. 1, p. 3396862, 2022.

[44] J. Franek and A. Kresta, "Judgment scales and consistency measure in AHP," Procedia economics and finance, vol. 12, pp. 164-173, 2014.

[45] B. D. Rouyendegh, A. Yildizbasi, and P. Üstünyer, "Intuitionistic fuzzy TOPSIS method for green supplier selection problem," Soft computing, vol. 24, no. 3, pp. 2215-2228, 2020.

[46] J. Y. Yap, C. C. Ho, and C.-Y. Ting, "Aggregating multiple decision makers' judgement," in Intelligent and Interactive Computing: Proceedings of IIC 2018: Springer, 2019, pp. 13-21.

[47] A. Naumann, S. Rieser, S. Musow, J. Hochweber, and J. Hartig, "Sensitivity of test items to teaching quality," Learning and Instruction, vol. 60, pp. 41-53, 2019.

[48] B. Gładysz and A. Kasperski, "Computing mean absolute deviation under uncertainty," Applied Soft Computing, vol. 10, no. 2, pp. 361-366, 2010.

[49] K. Elsayed, "Mean absolute deviation: analysis and applications," International Journal of Business and Statistical Analysis, vol. 2, no. 02, 2015.

[50] T. O. Hodson, T. M. Over, and S. S. Foks, "Mean squared error, deconstructed," Journal of Advances in Modeling Earth Systems, vol. 13, no. 12, p. e2021MS002681, 2021.

[51] G. W. Corder and D. I. Foreman, Nonparametric statistics: A step-by-step approach. John Wiley & Sons, 2014.

[52] Ž. Stević, M. Baydaş, M. Kavacık, E. Ayhan, and D. Marinković, "Selection of data conversion technique via sensitivity-performance matching: ranking of small e-vans with PROBID method," Facta Universitatis, Series: Mechanical Engineering, pp. 643-671, 2024.